

Enhancing the Reasoning Ability of Multimodal Large Language Models via Mixed Preference Optimization

Weiyun Wang^{2,1}, Zhe Chen^{3,1}, Wenhai Wang^{4,1}, Yue Cao^{3,1}, Yangzhou Liu^{3,1}, Zhangwei Gao¹, Jinguo Zhu¹, Xizhou Zhu^{5,1}, Lewei Lu⁶, Yu Qiao¹, Jifeng Dai^{5,1}✉

¹OpenGVLab, Shanghai AI Laboratory, ²Fudan University, ³Nanjing University, ⁴The Chinese University of Hong Kong, ⁵Tsinghua University, ⁶SenseTime Research

[Project Page](#)

Abstract

Existing open-source multimodal large language models (MLLMs) generally follow a training process involving pre-training and supervised fine-tuning. However, these models suffer from distribution shifts, which limit their multimodal reasoning, particularly in the Chain-of-Thought (CoT) performance. To address this, we introduce a preference optimization (PO) process to enhance the multimodal reasoning capabilities of MLLMs. Specifically, (1) on the data side, we design an automated preference data construction pipeline to create MMPR, a high-quality, large-scale multimodal reasoning preference dataset; and (2) on the model side, we explore integrating PO with MLLMs, developing a simple yet effective method, termed Mixed Preference Optimization (MPO), which boosts multimodal CoT performance. Our approach demonstrates improved performance across multiple benchmarks, particularly in multimodal reasoning tasks. Notably, our model, InternVL2-8B-MPO, achieves an accuracy of 67.0 on MathVista, outperforming InternVL2-8B by 8.7 points and achieving performance comparable to the 10× larger InternVL2-76B. We hope this study could inspire further advancements in MLLMs. Code, data, and model shall be publicly released.

1. Introduction

With the remarkable success of large language models (LLMs) [1, 5, 10, 11, 26, 89, 92, 93] in the field of natural language processing, the training paradigm comprising pre-training and supervised fine-tuning (SFT) have also swept the multimodal field, becoming the primary choice for the research and development of multimodal large language models (MLLMs). Benefiting from the large-scale

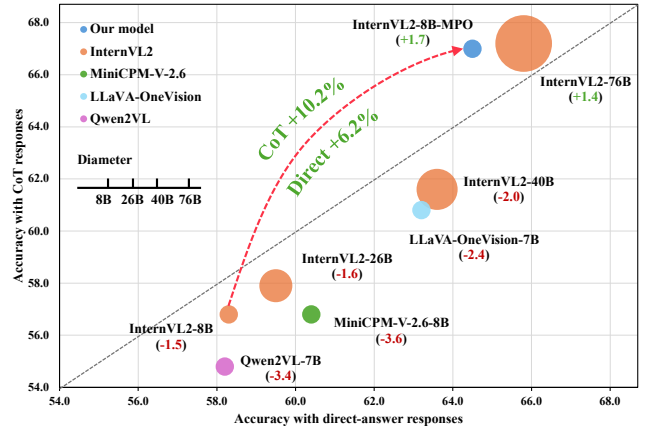


Figure 1. **Open-source model performance on MathVista.** The X- and Y-axes represent the accuracy evaluated with direct-answer responses and CoT responses, respectively. The bubble size is positively correlated with the number of model parameters. The values in parentheses indicate the performance gap between CoT and direct-answer responses. Notably, most open-source models perform worse when answering with CoT.

pre-training corpora [43, 48, 80, 90, 99, 114] and high-quality SFT data [20, 24, 53, 55, 98], a series of open-source MLLMs [6, 20, 44, 46, 52, 96, 98, 105] exhibit strong performance across various domain and tasks, some even achieving results comparable to commercial models such as GPT-4o [70] and Gemini [78, 88].

However, open-source MLLMs still exhibit limited reasoning capabilities. As shown in Figure 1, InternVL2-8B [20] achieves a score of 58.3 on MathVista [61], a benchmark for multimodal reasoning, when using direct answers but drops to 56.8 with Chain-of-Thought (CoT) reasoning, indicating that CoT reasoning actually reduces its performance. This decline is commonly observed across open-source MLLMs [20, 44, 96, 105]. We attribute this phenomenon primarily to a *distribution shift* introduced by the

✉ Corresponding Author: daijifeng@tsinghua.edu.cn

SFT loss. Specifically, SFT relies on teacher forcing, where the model is trained to predict the next token based on previous ground-truth tokens. However, during inference, models must predict each token based on their own prior outputs, leading to a distribution shift between training and inference. Since the direct-answer approach requires only brief responses, while CoT reasoning involves generating a long rationale, the distribution shift problem becomes more severe during CoT. This results in models performing worse with CoT reasoning compared to direct-answer responses.

To address the limitations of CoT reasoning in MLLMs, we draw inspiration from recent NLP approaches [42, 74, 103] that use Preference Optimization (PO) techniques to align model outputs with desired reasoning patterns. Specifically, methods like Direct Preference Optimization (DPO) [76] allow models to learn from preference signals to generate responses that better align with user requirements, offering the foundation for Reinforcement Learning from Human Feedback (RLHF). While RLHF has been explored for MLLMs primarily to reduce hallucinations [18, 85, 106], its application for enhancing multimodal reasoning remains under-explored. Building on these insights, we conduct a systematic study on using PO to strengthen the multimodal reasoning capabilities of MLLMs.

Enhancing the multimodal reasoning abilities of MLLMs through PO presents several challenges: (1) *Limited multimodal reasoning preference data and high annotation cost.* Existing multimodal preference datasets [47, 85, 106, 107, 111] primarily address hallucination issues and focus on natural images and perception data, lacking scientific images and reasoning data. Annotating these types of data requires human annotators to carefully compare the given reasoning processes, making it both time-consuming and costly. (2) *Lack of open-source methods for improving multimodal reasoning via PO.* Although previous works have explored fine-tuning MLLMs using feedback from various sources, these models typically exhibit performance gains on hallucination benchmarks, with little enhancement in general reasoning abilities. Thus, leveraging PO to improve multimodal reasoning capabilities remains largely under-explored.

This work addresses these challenges from both the data and model sides. (1) *On the data side*, we design an automated preference data construction pipeline to create MMPR, a high-quality, large-scale multimodal reasoning preference dataset. (2) *On the model side*, we explore various PO methods with MLLMs, introducing a simple yet effective method, termed Mixed Preference Optimization (MPO), which boosts multimodal CoT performance without the requirement for a reward model.

Specifically, we propose a continuation-based pipeline called Dropout Next Token Prediction (DropoutNTP) for samples lacking clear ground truth and a correctness-based

pipeline for samples with clear ground truth. In DropoutNTP, the responses generated by InternVL2-8B are considered as positive samples. For a given chosen response, we truncate it by half and then prompt InternVL2-8B to complete the remaining portion of the truncated answer without access to the image input. This generated completion serves as the rejected answer for the paired sample. Experimental results in Section 5.2 demonstrate that this straightforward method achieves comparable performance in reducing hallucinations compared to the divide-and-conquer method proposed in RLAI-F-V [107]. In the correctness-based pipeline, multiple solutions to each question are sampled from InternVL2-8B. Solutions matching the ground truth answer are used as chosen responses, while those that do not are used as rejected responses.

Additionally, we propose the MPO method. The key insight behind this algorithm is that *an effective PO process should enable the model to learn the relative preference between pairs of responses, the absolute quality of individual responses, and the process for generating preferred responses.* Compared to previous multimodal PO methods [47, 75, 85, 106, 107, 111], our approach excels in the following aspects: (1) *Efficient automated data construction pipeline:* Our pipeline enables high-quality preference pair generation at a controlled cost. (2) *Effectiveness across diverse domains:* Models fine-tuned with our data and approach show superior performance across reasoning, question-answering, and hallucination benchmarks. (3) *Improvements over SoTA settings:* Our results are based on InternVL2-8B, one of the leading open-source MLLMs, further highlighting the potential of our method.

In summary, our main contributions are as follows:

(1) We propose an efficient preference data construction pipeline. Based on this pipeline, we create MMPR, a high-quality, large-scale multimodal reasoning preference dataset containing approximately 3 million samples.

(2) We introduce MPO, an effective PO algorithm designed to improve the reasoning abilities of MLLMs. The resulting model, InternVL2-8B-MPO, exhibits enhanced multimodal reasoning ability and fewer hallucinations compared to its baseline model (*i.e.*, InternVL2-8B).

(3) We conduct extensive experiments to explore practical approaches for improving multimodal reasoning via PO. Results show that PO significantly improves reasoning abilities over SFT. Notably, the proposed InternVL2-8B-MPO achieves an accuracy of 67.0 on MathVista [61], outperforming InternVL2-8B by 8.7 points and achieving performance comparable to the $10\times$ larger InternVL2-76B.

2. Related Work

Multimodal Large Language Models. With advancements in LLMs, significant progress has also been made in MLLMs. To leverage the abilities of pre-trained LLMs [5,

11, 26] and Vision Foundation Models (VFM) [19, 77], a series of works [20, 45, 46, 53, 56, 96, 99, 100] employ a connector to align their latent space, achieving promising performance at a controllable cost. Besides, another series of works [2, 26, 91, 97] extend pre-trained LLMs with additional fusion layers for vision features, reducing the number of visual tokens required by LLMs while introducing extra training costs. Recently, there have been explorations into vision encoder-free architectures [7, 50, 62, 87, 101], which consists of a single transformer model that jointly processes visual and textual information without a separate encoder. In addition to exploring model architectures, recent works [27, 48, 55, 98, 104, 109] also try to construct high-quality training data to improve multimodal reasoning abilities. Despite these advancements, MLLMs typically rely on a training paradigm comprising pre-training and supervised fine-tuning, which suffers from the curve of distribution shift and exhibits limited multimodal reasoning abilities. In this work, we conduct a systematic study on using preference optimization to enhance the multimodal reasoning ability of MLLMs.

Preference Optimization. Preference optimization (PO) is a crucial technique for advancing LLMs and MLLMs. Specifically, Reinforcement Learning from Human Feedback (RLHF) uses human preferences as a reward signal to fine-tune models, aligning them with human preferences. InstructGPT [72] employs a reward model as a proxy for human preferences and maximizes this reward via the PPO algorithm [81], improving the model’s ability to follow user intent and become more helpful, honest, and harmless (3H). PPO-Max [94, 112] carefully explores the implementation details of PPO, proposing a more stable version of the algorithm. Additionally, DPO [76] proposes an efficient PO algorithm based on the Bradley-Terry model [9], removing the need for an explicit reward model. Subsequent works [4, 21, 25, 28, 32, 42, 54] have further analyzed and refined this method from various perspectives. In natural language processing, a series of works [42, 74] have explored how to leverage PO to enhance reasoning ability. In the multimodal field, however, most methods [47, 85, 106, 107, 111] primarily focus on reducing hallucination, leaving the potential for PO to improve multimodal reasoning ability under-explored. This work demonstrates that PO not only mitigates hallucinations but also strengthens multimodal reasoning abilities, highlighting its broader applicability in MLLM development.

3. Scalable Multimodal Preference Dataset Generation

To address the scarcity of multimodal preference data, we introduce a scalable data construction pipeline. Based on this pipeline, we construct a million-level MultiModal PReference dataset (MMPR).

3.1. Data Engine

Definition. Each data sample in our MMPR consists of an image $I \in \mathcal{I}$, an instruction $x \in \mathcal{X}$, a chosen response $y_c \in \mathcal{Y}_p$, and a rejected response $y_r \in \mathcal{Y}_n$, where y_c is preferable to y_r . The image sets \mathcal{I} and instruction sets \mathcal{X} are collected from existing datasets. \mathcal{Y}_p and \mathcal{Y}_n represent the positive and negative response set, respectively. Given a certain image I and instruction x , we sample the candidate response y from an initial instruction model M_0 as follows:

$$y \sim M_0(y | x, I), \quad (1)$$

where $M_0(y | x, I)$ represents the response distribution of M_0 conditioned on image I and instruction x .

For instructions with clear ground truths, the model is prompted to first provide the reasoning process and then give the final answer in the format like “Final Answer: ***”. Responses matching the ground truth answer constitute the positive set \mathcal{Y}_p , while those that do not match make up the negative set \mathcal{Y}_n . Additionally, responses that fail to provide a clear final answer are also merged into \mathcal{Y}_n . Given these responses labeled as positive or negative, we build the preference pairs by selecting a chosen response y_c from \mathcal{Y}_p and a negative response y_r from \mathcal{Y}_n .

For instructions without clear ground truths, we propose a simple yet effective method: Dropout Next-Token Prediction (Dropout NTP). Specifically, we directly consider all responses generated from equation 1 as positive set \mathcal{Y}_p . To generate the negative set \mathcal{Y}_n , we sample a response y from \mathcal{Y}_p and drop the last half of this response. The model is required to complete the remained response as follows:

$$\tilde{y}_{\geq j} \sim M_0(\tilde{y}_{\geq j} | x, y_{< j}), \quad (2)$$

where $y_{< j}$ and $y_{\geq j}$ is the remained part and truncated part of y , respectively. $\tilde{y}_{\geq j}$ is the completion of $y_{< j}$ without the image input. The original response $y = [y_{< j}, y_{\geq j}]$ serves as the chosen response y_c and the completed response $\tilde{y} = [y_{< j}, \tilde{y}_{\geq j}]$ serves as the rejected response y_r . It is worth noting that while the responses generated by M_0 may not be perfect, the completions generated without the image input will introduce more hallucinations than those generated with the image input. Therefore, the partial order relationship between y and \tilde{y} holds true.

Compared with previous methods, our data engine is as effective as the more complex divide-and-conquer method proposed in RLAI-F-V [107] (see the experimental results in Section 5.2.2), while more efficient. Taking data generation for M3CoT as an example, our pipeline incurs a token cost of 571.2 per preference pair, compared to 992.7 tokens for the divide-and-conquer approach used in RLAI-F-V. Thus, the cost of our pipeline is only 57.5% of that of RLAI-F-V.

Instruction with clear ground truths



Question:

Your task is to answer the question below. Give step by step reasoning before you answer, and when you're ready to answer, please use the format "Final answer: ..."

Question:

What could be the reason for a person to have a brown teddy bear tied with a ribbon around its neck, a pack of colored pencils and a colorful patterned blanket?

- A. The person likes to collect teddy bears, color pencils, and blankets as they are a fan of art
- B. The person received the brown teddy bear as a gift from someone and owns the colored pencils because they are a fan of art
- C. The person is a child who loves to play with teddy bears and draw with color pencils
- D. The person is a school teacher who uses these items to decorate their classroom for their students

Chosen Response:

To determine the answer, let's analyze each option:

- A. The person likes to collect teddy bears, color pencils, and blankets as they are a fan of art. This option assumes that the person collects multiple items as part of a fan interest in art. The items are not just limited to teddy bears, colored pencils, and blankets.
- B. The person received the brown teddy bear as a gift from someone and owns the colored pencils because they are a fan of art. This option suggests that the teddy bear was received as a gift and that the person uses colored pencils as they are a fan of art. This explanation is plausible.
- C. The person is a child who loves to play with teddy bears and draw with color pencils. This option simply states that the person is a child, but it does not provide a clear connection between the other items like the vibrant blanket.
- D. The person is a school teacher who uses these items to decorate their classroom for their students. This option seems less likely because the items (teddy bear, colored pencils, and blanket) might not be common classroom decoration items.

Final answer: B

Rejected Response:

To determine the reason for a person to have a brown teddy bear tied with a ribbon around its neck, a pack of colored pencils, and a colorful patterned blanket, we need to consider the context and meaning behind these items.

- Option A: "The person likes to collect teddy bears, color pencils, and blankets as they are a fan of art." This option suggests that the person has a specific interest in collecting items, which may not necessarily imply a connection to art.
- Option B: "The person received the brown teddy bear as a gift from someone and owns the colored pencils because they are a fan of art." This option suggests a personal connection to a gift while considering an individual interest in art.
- Option C: "The person is a child who loves to play with teddy bears and draw with color pencils." **This option assumes that the person is a child, which could be a plausible explanation for the items.**
- Option D: "The person is a school teacher who uses these items to decorate their classroom for their students." This option suggests a professional context rather than personal interest or collecting items.

Given the context of the items in the image and the most plausible explanation based on the presence of children's items and a child-friendly material like a teddy bear with a ribbon, the most likely explanation is:

Final answer: C

Instruction without clear ground truths



Question:

Where is the woman in the picture located?

Chosen Response:

The woman in the picture is located in a public park. Here are the observations that support this conclusion:

- 1. Surroundings: The image shows a well-maintained area with grass, trees, and benches, which is typical of public parks. The presence of other people walking in the background also suggests a public space.
- 2. Lighting and Ambience: The bright daylight and open space are characteristic of outdoor environments like parks, rather than indoor locations like libraries.
- 3. Bench Design: The style of the bench, which appears to be made of wood or metal with a decorative back, is common in public parks. Such benches are often placed along walking paths for public use.
- 4. Vegetation: The variety and arrangement of trees and shrubs look like planned landscaping, which is typical of parks rather than the natural, untamed growth found in a virgin forest. These elements collectively indicate that the setting is a public park.

Rejected Response:

The woman in the picture is located in a public park. Here are the observations that support this conclusion:

- 1. Surroundings: The image shows a well-maintained area with grass, trees, and benches, which is typical of public parks. The presence of other people walking in the background also suggests a public space.
- 2. Lighting and Ambience: The bright daylight and open space are characteristic of outdoor environments like parks, rather than indoor locations like libraries.
- 3. Bench: If she were in a private garden or library, it would likely be more private and less likely to have communal spaces such as a bench.
- 4. Background Activity: **The presence of other people in the background suggests a public space where people gather, such as a park.** Given these observations, the picture most likely shows a woman in a public park.

Figure 2. **Data examples in MMRP.** For instructions with clear ground truths, we propose a correctness-based pipeline, which samples multiple solutions and considers those with correct answers as chosen responses and those with incorrect answers as rejected responses. For instructions without clear ground truths, we propose DropoutNTP to generate rejected responses. Differences between the chosen and rejected responses are emphasized in *italicized text*. **Red** highlights incorrect responses.

3.2. Multimodal Preference Dataset

Dataset Statistics. Using this pipeline, we build a large-scale multimodal preference dataset, MMRP. Data examples are presented in Figure 2. See more examples in the Appendix. This dataset comprises approximately 750K samples without clear ground truths and 2.5M samples with clear ground truths. For samples without clear ground truths, each instruction averages 25.0 tokens, while the chosen and rejected responses average 211.4 and 171.2 tokens, respectively. The longest chosen and rejected responses consist of 1,342 and 1,642 tokens, respectively, whereas the shortest chosen and rejected responses contain 20 and 17 tokens, respectively. For samples with clear ground truths, the average instruction length is 79.5 tokens, with the chosen and rejected responses averaging 300.0 and 350.5 tokens, respectively. The longest chosen and rejected responses are composed of 2,018 and 4,097 tokens, while the shortest responses contain 32 and 33 tokens, respectively.

Data Source. As shown in Table 1, to ensure the diversity of instructions and images, we collect samples from diverse domains, including general visual question answering (VQA) [29, 34, 59, 63], science [16, 39, 60], chart [13, 37, 64], mathematics [12, 27, 38, 51, 58, 82], OCR [8, 33, 66, 68, 83], and document [22]. Notably, when constructing open-ended samples, we collect instructions

Task	Dataset
General VQA	VQAv2 [29], GQA [34], OKVQA [63], IconQA [59]
Science	AI2D [39], ScienceQA [60], M3CoT [16]
Chart	ChartQA [64], DVQA [37], MapQA [13]
Mathematics	GeoQA+ [12], CLEVR-Math [51], Geometry3K [58], GEOS [82], GeomVerse [38], Geo170K [27]
OCR	OCRVQA [68], InfoVQA [66], TextVQA [83], STVQA [8], SROIE [33]
Document	DocVQA [65]

Table 1. **Datasets used to build our preference dataset.** We collect images and instructions from various tasks to ensure the diversity of our dataset.

from all the data sources mentioned above and prompt the model to answer the original question without additional requirements. On the other side, when building samples through the correctness-based pipeline, we exclude questions from general VQA and document sources, as verifying the correctness of the generated answers using heuristic rules is challenging for datasets in these domains. For example, the ground truths in VQAv2 [29] consist of a single word or phrase, which may lead to false-negative responses when the model outputs a complete sentence or a synonym as the final answer. Such false-negative responses can negatively impact training effectiveness.

4. Improved Multimodal Large Language Model with Preference Optimization

To enhance the multimodal reasoning capabilities of MLLMs, we propose mixed preference optimization (MPO), a method that blends supervised fine-tuning (SFT) loss with various preference optimization losses to enhance training effectiveness. Additionally, we investigate different Chain-of-Thought (CoT) approaches with multimodal input to improve reasoning performance.

4.1. Mixed Preference Optimization

We observed that when MLLMs are trained on large-scale preference datasets using direct preference optimization (DPO), they might fail to generate reasonable rationales and produce gibberish. This phenomenon aligns with the analysis presented in Smaug [73]. To address this issue, we introduce the MPO in this work, aiming to learn the relative preference between pairs of responses, the absolute quality of individual responses, and the process for generating preferred responses.

Training Objective. MPO is defined as a combination of preference loss \mathcal{L}_p , quality loss \mathcal{L}_q , and generation loss \mathcal{L}_g , which can be formulated as follows:

$$\mathcal{L} = w_p \mathcal{L}_p + w_q \mathcal{L}_q + w_g \mathcal{L}_g, \quad (3)$$

where w_* represents the weight assigned to each loss component. In this work, we empirically compare different variants of preference loss [4, 14, 21, 32, 36, 54, 67, 69, 76, 102]. Based on the experimental results, we use DPO [76] as our preference loss and BCO [36] as our quality loss.

Preference Loss. The DPO [76] serves as the preference loss to enable the model to learn the relative preference between chosen and rejected responses. DPO eliminates the requirement of training an explicit reward model based on the assumption of the Bradley-Terry model [9] and optimizes the following loss function:

$$\mathcal{L}_p = -\log \sigma \left(\beta \log \frac{\pi_\theta(y_c | x)}{\pi_0(y_c | x)} - \beta \log \frac{\pi_\theta(y_r | x)}{\pi_0(y_r | x)} \right), \quad (4)$$

where β is the KL penalty coefficient, and x , y_c , and y_r are user query, chosen response, and rejected response, respectively. The policy model π_θ is initialized from model π_0 .

Quality Loss. The BCO loss [36] is employed as the quality loss, which helps the model to understand the absolute quality of individual responses. This algorithm trains a binary classifier, where the logit serves as a reward and effectively maps the chosen response to 1 and the rejected response to 0. The loss function is defined as:

$$\mathcal{L}_q = \mathcal{L}_q^+ + \mathcal{L}_q^-, \quad (5)$$

where \mathcal{L}_q^+ and \mathcal{L}_q^- represent the loss for chosen and rejected responses, respectively. They are calculated independently,

requiring the model to differentiate the absolute quality of individual responses. The loss terms are given by:

$$\mathcal{L}_q^+ = -\log \sigma \left(\beta \log \frac{\pi_\theta(y_c | x)}{\pi_0(y_c | x)} - \delta \right), \quad (6)$$

$$\mathcal{L}_q^- = -\log \sigma \left(-\left(\beta \log \frac{\pi_\theta(y_r | x)}{\pi_0(y_r | x)} - \delta \right) \right), \quad (7)$$

where δ represents the reward shift, calculated as the moving average of previous rewards to stabilize training.

Generation Loss. The SFT loss is used as the generation loss to help the model learn the generation process of preferred responses. The loss function is defined as:

$$\mathcal{L}_g = -\frac{\log \pi_\theta(y_c | x)}{|y_c|}. \quad (8)$$

4.2. Chain-of-Thought with Multimodal Input

During the data sampling process, we require the model to provide a detailed CoT reasoning process instead of directly answering the final answer. For most samples, we sample the responses using the prompt shown in the bottom case of Figure 2, which requires the model to perform a step-by-step analysis. Considering that multimodal models involve non-textual inputs, we further introduce the following CoT methods: (1) **Background Knowledge-based CoT:** The model first introduces relevant background knowledge related to the problem or image, followed by reasoning steps and the final answer. This approach is applied to samples from the science domain. (2) **Visual Content-based CoT:** The model begins by analyzing the visual contents in the image, then proceeds with reasoning and the final answer. This method is used for samples from chart, OCR, and document domains. (3) **Grounded CoT:** The model generates a text response while simultaneously linking all referenced objects in the response to corresponding regions in the image. This approach is applied to general VQA domain samples. Responses generated by these above CoT methods are mixed with those sampled using the prompt shown in the bottom case of Figure 2. These approaches not only effectively integrate multimodal information into the reasoning process but also enhance data diversity. Furthermore, including the background knowledge and visual contents at the start of responses also improves the quality of the negative responses generated by DropoutNTP, preventing a significant quality gap between positive and negative samples that could reduce training effectiveness.

5. Experiments

5.1. Main Results

In this section, we compare our InternVL2-8B-MPO with leading MLLMs on multimodal reasoning [16, 61, 95],

Model Name	Reasoning			General VQA		Hallucination Evaluation		
	M3CoT	MathVista	MathVision	MMVet	LLaVA-Bench	POPE	CRPE	MMHalBench
<i>Closed-Source Models</i>								
Gemini-1.5-Pro [78]	-	63.9	19.2	-	-	-	-	-
GPT-4o [71]	64.3	63.8	30.4	69.1	97.6	86.9	76.6	4.0
GPT-4o-Mini [71]	61.9	52.4	27.3	66.9	95.4	85.1	73.1	3.6
<i>Open-Source Models</i>								
LLaVA-1.5-13B [52]	39.5	27.6	11.1	36.3	70.7	85.9	55.6	2.4
Qwen2-VL-7B [96]	57.8	58.2	21.1	60.6	67.7	88.1	74.4	3.4
MiniCPM-V-2-6-8B [105]	56.0	60.6	23.4	57.4	83.4	87.3	75.2	3.6
LLaVA-OneVision-7B [44]	52.3	63.2	18.4	51.4	79.9	88.4	73.7	3.1
<i>InternVL Models</i>								
InternVL2-26B [20]	58.2	59.4	23.4	62.1	92.3	88.0	75.6	3.7
InternVL2-40B [20]	63.6	63.7	21.4	65.5	100.5	88.4	77.3	3.9
InternVL2-76B [20]	65.4	67.2	23.7	65.7	99.3	89.0	77.8	3.8
InternVL2-Pro [20]	65.6	66.3	18.8	69.4	99.5	88.2	77.6	3.7
InternVL2-8B [20]	59.3	58.3	20.4	54.2	73.2	86.9	75.0	3.3
InternVL2-8B-MPO (ours)	79.2	67.0	25.7	56.2	76.7	88.1	75.4	3.5

Table 2. **Results on 8 multimodal benchmarks.** We report the overall score of MM-Vet and LLaVA-Bench evaluated by GPT-4-Turbo. Our InternVL2-8B-MPO demonstrates superior performance compared to InternVL2-8B across multimodal reasoning, VQA, and hallucination evaluation benchmarks. Notably, our model even achieves reasoning performance comparable to the 10× larger InternVL2-76B.

complex Visual Question Answering (VQA) [53, 108], and hallucination evaluation [49, 85, 98] tasks.

Benchmarks. For the multimodal reasoning task, we evaluate our model on three benchmarks: (1) **M3CoT** [16], a comprehensive benchmark designed to evaluate the multimodal CoT reasoning abilities of models. (2) **MathVista** [61], a widely-used benchmark for evaluating multimodal mathematical reasoning capabilities. (3) **MathVision** [95], which collects evaluation data from real math competitions and presents a greater challenge compared to MathVista. We report accuracy for these benchmarks.

For the complex VQA task, we evaluate our model on two benchmarks: (1) **MM-Vet** [108], which evaluates the model’s ability to engage in visual conversations across a diverse range of tasks. (2) **LLaVA-Bench** [53], a commonly-used benchmark for assessing multimodal conversation, detailed description, and complex reasoning capabilities with open-ended questions. Both benchmarks use GPT-4 to evaluate the correctness and helpfulness of responses. We report the overall score for these benchmarks.

For the hallucination evaluation task, we evaluate our model on three benchmarks: (1) **POPE** [49], which measures the hallucination level of object existence using Yes/No questions. We report the F1 score for this benchmark. (2) **CRPE** [98], which measures the hallucination level of the relation between objects using multiple-choice questions. We report accuracy for this benchmark. (3) **MMHal-Bench** [85], which consists of open-ended questions where GPT-4 compares model outputs to human responses, assessing hallucination rate and informativeness.

We report the overall score for this benchmark.

Results. As shown in Table 2, our InternVL2-8B-MPO achieves superior performance across all benchmarks, particularly excelling in multimodal reasoning tasks. On the MathVista benchmark, our model achieves an accuracy of 67.0%, outperforming InternVL2-8B by 8.7 points and achieving performance comparable to the 10× larger InternVL2-76B. On the MathVision benchmark, our model achieves an accuracy of 25.7%, establishing a new state-of-the-art performance among open-source models. These results demonstrate the effectiveness of our preference optimization approach in enhancing multimodal reasoning capabilities. Additionally, on the POPE benchmark, our model exhibits a 1.2-point improvement over InternVL2-8B, demonstrating the effectiveness of the perception data contained in our MMPR dataset to mitigate hallucinations. Furthermore, our model also shows superior performance compared to the InternVL2-8B on complex VQA benchmarks, indicating that the general abilities of our model are also improved, benefiting from enhanced reasoning abilities and mitigated hallucinations.

5.2. Ablation Study

In this section, we present ablation studies to analyze the effects of preference optimization and SFT on multimodal reasoning abilities. Additionally, we compare our proposed DropoutNTP method with the divide-and-conquer approach from RLAIIF-V [107], demonstrating the effectiveness of our approach. Furthermore, we conduct extensive experiments to analyze the effects of different preference opti-

Model Name	Setting	M3CoT	MathVista	MMVet	POPE
InternVL2-8B	Direct	59.3	58.3	54.2	86.9
	CoT	57.0	56.8	54.7	82.9
InternVL2-8B-SFT	Direct	63.9	62.7	54.7	86.5
	CoT	67.8	64.2	53.8	84.0
InternVL2-8B-MPO	Direct	77.2	64.5	55.1	87.0
	CoT	79.2	67.0	56.2	88.1

Table 3. **Results of models trained with SFT and MPO.** The SFT training data consists of the chosen responses from the preference pairs used in MPO training. In the Direct setting, the model is prompted to provide the answer directly, while in the CoT setting, the model is instructed to answer with detailed rationales.

mization algorithms. We also present analysis of the effects on text-only performance.

5.2.1. Comparison between MPO and SFT

To compare the impact of MPO and SFT on improving multimodal reasoning ability, we use the chosen responses in MMPr as SFT data to fine-tune InternVL2-8B. As shown in Table 3, the results indicate that the model trained with MPO consistently outperforms that trained with SFT across all benchmarks. For example, the MPO-trained model achieves a score of 79.2 on the multimodal reasoning benchmark M3CoT, surpassing its SFT counterpart by 11.4 points. Furthermore, the MPO-trained model also performs better on the general benchmark (MMVet) and the hallucination benchmark (POPE). Notably, the SFT-trained model performs worse with CoT responses than with direct-answer responses on MMVet and POPE, demonstrating that SFT alone is insufficient to enhance multimodal CoT abilities. These results demonstrate that while SFT provides moderate improvement, preference optimization is more effective in improving the overall performance of the model.

5.2.2. Comparison with RLAI-F-V

Here, we compare our proposed Dropout Next-Token Prediction (Dropout NTP) method with the divide-and-conquer approach from RLAI-F-V [107]. To ensure a fair comparison, we use the same prompts and chosen responses as in RLAI-F-V and replace the rejected responses with those generated by continuation without image input. Following RLAI-F-V, we report the hallucination rates in response-level (Resp.) and mention-level (Ment.) for Object HalBench [79] and overall score and hallucination rates (Hall.) for MMHal-Bench [85]. As shown in Table 4, the model trained with our data achieves performance comparable to that of the model trained with RLAI-F-V, demonstrating the effectiveness of our method. Specifically, the response-level hallucination rate of the model trained with our data on Object HalBench is 7.6, compared to 7.3 for its counterpart. Besides, this model achieves a score of 3.6 on the MMHal-Bench, compared to 3.5 for its counterpart. Note that our

Method	Object HalBench		MM HalBench	
	Resp. (↓)	Ment. (↓)	Score	Hall. (↓)
InternVL2-8B	18.4	8.7	3.3	40.6
RLAI-F-V [107]	7.3	3.9	3.5	32.3
DropoutNTP (ours)	7.6	4.1	3.6	31.3

Table 4. **Comparison of DropoutNTP and the divide-and-conquer approach from RLAI-F-V.** We replace negative samples in RLAI-F-V with the responses generated using DropoutNTP.

method requires the model to generate only a single continuation for each sample, while RLAI-F-V requires the model to decompose the response into atomic claims and then verify each one individually. Therefore, our method is more efficient. A quantitative analysis is provided in Section 3.1.

5.2.3. Effects of optimization algorithms

Here, we empirically compare the effectiveness of different optimization algorithms, including (1) **DPO** [76], which directly fine-tunes the model on an offline preference dataset without explicitly constructing a reward function. (2) **RSO** [54], which applies a hinge loss on the normalized likelihood instead of the sigmoid loss used in DPO. (3) **IPO** [4], which introduces a modified loss function to address overfitting in DPO by averaging log-likelihoods and controlling the gap between chosen and rejected completions via a beta parameter. (4) **cdPO** [69], which is a modification of the DPO loss that accounts for potential label noise in preference data. (5) **RobustDPO** [21], which provides an unbiased estimate of the DPO loss designed to handle preference noise in data. Similar to cdPO, it assumes that labels are noisy with a certain probability. (6) **BCO** [36], which introduces a binary classifier trained to output logits used as reward values. (7) **SPPO** [102], which iteratively pushes chosen rewards toward 1/2 and rejected rewards toward -1/2 to approximate a Nash equilibrium, aiming to reduce data sparsity issues. (8) **AOT** [67], which applies Distributional Preference Alignment via Optimal Transport. (9) **TR-DPO** [28], which adds synchronization between the model and a reference model every few steps to mitigate overfitting during DPO training. (10) **ORPO** [32], a reference model-free preference optimization algorithm that uses a log odds ratio penalty appended to the NLL loss, allowing for preference-aligned fine-tuning without an additional preference alignment phase. For all algorithms, we set the learning rate to $5e-6$ and use the hyper-parameters suggested in their corresponding paper. Additionally, we extend these algorithms with SFT loss to analyze its impact. The SFT model trained with the chosen responses of the reasoning preference data is also included as a baseline.

Notably, most current benchmarks lack corresponding in-distribution training samples, and the data distribution of our MMPr may differ from that of these benchmarks. This

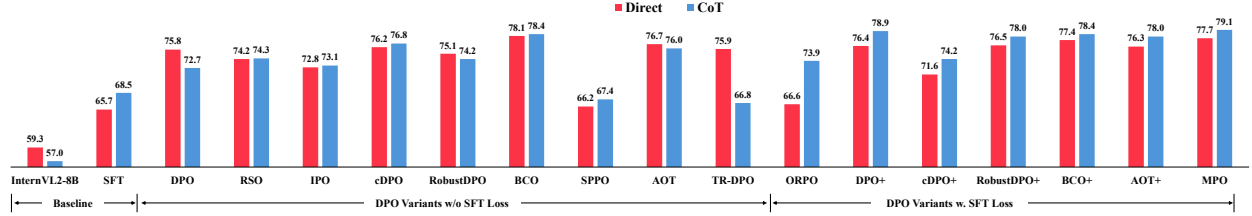


Figure 3. **Results of models trained with different preference optimization algorithms on M3CoT.** The algorithm X extended with the SFT loss is called X+ for brevity. For instance, DPO+ denotes the combination of DPO loss and SFT loss.

Setting	MMLU	Gaokao	TriviaQA	NQ	C3	Race-h	BBH	GSM8K	Math	TheoremQA	IFEval	HumanEval	MBPP	Average
Baseline	73.2	75.0	62.0	28.1	94.2	90.8	72.7	75.6	39.5	15.6	52.3	69.5	58.8	62.1
SFT	71.8	74.4	63.7	28.2	94.3	90.6	72.1	75.5	40.1	15.8	53.6	68.3	58.0	62.0
MPO	71.0	74.8	64.2	29.3	94.2	90.6	71.8	75.0	40.4	20.8	56.4	68.9	61.5	63.0

Table 5. **Results on text-only benchmarks.** The model fine-tuned through MPO exhibits superior overall performance on text-only tasks compared to the baseline model and its SFT counterpart, particularly on TheoremQA and IFEval.

discrepancy can introduce additional variability when analyzing the impact of different optimization algorithms on training results. Therefore, we use the training and validation sets of M3CoT [16] for ablation studies.

The visualization results are illustrated in Figure 3 and the numerical results are presented in Table 6 and 7. We can observe that almost all preference optimization methods outperform their SFT counterpart in both the Direct and CoT settings. However, DPO and its variants struggle to enhance the CoT reasoning abilities of the model as the resulting models exhibit trivial or no improvement when answering with CoT reasoning responses compared to direct-answer responses. On the other hand, when combining SFT Loss with these DPO variants, all algorithms are able to improve the model’s CoT reasoning abilities, demonstrating that *the SFT loss is a key component for enhancing CoT reasoning abilities*. Additionally, models trained with TR-DPO, a DPO variant that updates the reference model every few steps, perform much worse when using CoT reasoning compared to direct-answer responses. Similarly, the model trained with ODPO, a reference-model-free method, achieves worse overall performance compared to other methods extended with SFT Loss. These results indicate that *the reference model constraint on policy updates is crucial for enhancing overall reasoning abilities, and the reference model should remain frozen during training*. Notably, models trained with DPO+ and BCO+ exhibit the best CoT performance among existing algorithms. Therefore, we use DPO and BCO as the preference loss and quality loss. The resulting algorithm (*i.e.*, MPO) further improves the overall performance.

5.3. Effects on text-only performance

We evaluate the text-only performance of our models on a series of benchmarks [3, 15, 17, 23, 30, 31, 35, 40, 41, 84,

86, 110, 113] and report the average performance across them. As shown in Table 5, although our MMR dataset does not include any text-only data, the MPO-trained model achieves superior average performance on these benchmarks compared to the baseline model. The most significant improvements are observed on TheoremQA and IFEval. Specifically, our model trained with MPO achieves an accuracy of 20.8 on TheoremQA, a benchmark consisting of complex science problems, outperforming the baseline model by 5.2 points and the SFT counterpart by 5.0 points. Additionally, since our dataset considers responses that fail to follow instructions as negative samples when constructing data using our correctness-based pipeline, our model also exhibits enhanced instruction-following abilities on IFEval, outperforming the baseline model by 4.1 points and the SFT counterpart by 2.8 points.

6. Conclusion

In this work, we introduce a preference optimization (PO) process to enhance the multimodal reasoning capabilities of MLLMs. On the data side, we design an automated pipeline for preference data construction, which is applicable to instructions both with and without clear ground truths. Using this pipeline, we create MMR, a high-quality, large-scale multimodal reasoning preference dataset. On the model side, we propose a simple yet effective method called Mixed Preference Optimization (MPO). This algorithm aims to learn the relative preference between pairs of responses, the absolute quality of individual responses, and the process for generating preferred responses. The resulting model, InternVL2-8B-MPO, exhibits enhanced multimodal reasoning ability and fewer hallucinations compared to its baseline model (*i.e.*, InternVL2-8B). We hope this study could inspire further advancements in MLLMs.

References

- [1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 1
- [2] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *NIPS*, 35:23716–23736, 2022. 3
- [3] Jacob Austin, Augustus Odena, Maxwell Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie Cai, Michael Terry, Quoc Le, et al. Program synthesis with large language models. *arXiv preprint arXiv:2108.07732*, 2021. 8
- [4] Mohammad Gheshlaghi Azar, Zhaohan Daniel Guo, Bilal Piot, Remi Munos, Mark Rowland, Michal Valko, and Daniele Calandriello. A general theoretical paradigm to understand learning from human preferences. In *International Conference on Artificial Intelligence and Statistics*, pages 4447–4455. PMLR, 2024. 3, 5, 7, 1
- [5] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023. 1, 2
- [6] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 2023. 1
- [7] Rohan Bavishi, Erich Elsen, Curtis Hawthorne, Maxwell Nye, Augustus Odena, Arushi Somani, and Saġnak Taşlılar. Introducing our multimodal models, 2023. 3
- [8] Ali Furkan Biten, Ruben Tito, Andres Mafla, Lluís Gomez, Marçal Rusinol, Ernest Valveny, CV Jawahar, and Dimosthenis Karatzas. Scene text visual question answering. In *ICCV*, pages 4291–4301, 2019. 4
- [9] Ralph Allan Bradley and Milton E Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952. 3, 5
- [10] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *NIPS*, 2020. 1
- [11] Zheng Cai, Maosong Cao, Haojiong Chen, Kai Chen, Keyu Chen, Xin Chen, Xun Chen, Zehui Chen, Zhi Chen, Pei Chu, et al. Internlm2 technical report. *arXiv preprint arXiv:2403.17297*, 2024. 1, 3
- [12] Jie Cao and Jing Xiao. An augmented benchmark dataset for geometric question answering through dual parallel text encoding. In *COLING*, pages 1511–1520, 2022. 4
- [13] Shuaichen Chang, David Palzer, Jialin Li, Eric Fosler-Lussier, and Ningchuan Xiao. Mapqa: A dataset for question answering on choropleth maps. *arXiv preprint arXiv:2211.08545*, 2022. 4
- [14] Huayu Chen, Guande He, Hang Su, and Jun Zhu. Noise contrastive alignment of language models with explicit rewards. *arXiv preprint arXiv:2402.05369*, 2024. 5
- [15] Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde De Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021. 8
- [16] Qiguang Chen, Libo Qin, Jin Zhang, Zhi Chen, Xiao Xu, and Wanxiang Che. M3cot: A novel benchmark for multi-domain multi-step multi-modal chain-of-thought. *arXiv preprint arXiv:2405.16473*, 2024. 4, 5, 6, 8, 2
- [17] Wenhui Chen, Ming Yin, Max Ku, Pan Lu, Yixin Wan, Xueguang Ma, Jianyu Xu, Xinyi Wang, and Tony Xia. Theoremqa: A theorem-driven question answering dataset. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7889–7901, 2023. 8
- [18] Yangyi Chen, Karan Sikka, Michael Cogswell, Heng Ji, and Ajay Divakaran. Dress: Instructing large vision-language models to align and interact with humans via natural language feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14239–14250, 2024. 2
- [19] Zhe Chen, Jiannan Wu, Wenhui Wang, Weijie Su, Guo Chen, Sen Xing, Zhong Muyan, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. *arXiv preprint arXiv:2312.14238*, 2023. 3
- [20] Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *arXiv preprint arXiv:2404.16821*, 2024. 1, 3, 6
- [21] Sayak Ray Chowdhury, Anush Kini, and Nagarajan Natarajan. Provably robust dpo: Aligning language models with noisy feedback. *arXiv preprint arXiv:2403.00409*, 2024. 3, 5, 7, 1
- [22] Christopher Clark and Matt Gardner. Simple and effective multi-paragraph reading comprehension. In *ACL*, pages 845–855, 2018. 4
- [23] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021. 8
- [24] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *NIPS*, 36, 2024. 1
- [25] Hanze Dong, Wei Xiong, Bo Pang, Haoxiang Wang, Han Zhao, Yingbo Zhou, Nan Jiang, Doyen Sahoo, Caiming Xiong, and Tong Zhang. Rlh workflow: From reward modeling to online rlhf. *arXiv preprint arXiv:2405.07863*, 2024. 3

- [26] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. 1, 3
- [27] Jiahui Gao, Renjie Pi, Jipeng Zhang, Jiacheng Ye, Wan-jun Zhong, Yufei Wang, Lanqing Hong, Jianhua Han, Hang Xu, Zhenguo Li, et al. G-llava: Solving geometric problem with multi-modal large language model. *arXiv preprint arXiv:2312.11370*, 2023. 3, 4
- [28] Alexey Gorbatoyski, Boris Shaposhnikov, Alexey Malakhov, Nikita Surnachev, Yaroslav Aksenov, Ian Maksimov, Nikita Balagansky, and Daniil Gavrilov. Learn your reference model for real good alignment. *arXiv preprint arXiv:2404.09656*, 2024. 3, 7, 1
- [29] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *CVPR*, pages 6904–6913, 2017. 4
- [30] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2020. 8
- [31] Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*, 2021. 8
- [32] Jiwoo Hong, Noah Lee, and James Thorne. Orpo: Monolithic preference optimization without reference model. *arXiv preprint arXiv:2403.07691*, 2(4):5, 2024. 3, 5, 7, 1
- [33] Zheng Huang, Kai Chen, Jianhua He, Xiang Bai, Dimosthenis Karatzas, Shijian Lu, and CV Jawahar. Icdar2019 competition on scanned receipt ocr and information extraction. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 1516–1520. IEEE, 2019. 4
- [34] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *CVPR*, pages 6700–6709, 2019. 4
- [35] Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. *arXiv preprint arXiv:1705.03551*, 2017. 8
- [36] Seungjae Jung, Gunsoo Han, Daniel Wontae Nam, and Kyoung-Woon On. Binary classifier optimization for large language model alignment. *arXiv preprint arXiv:2404.04656*, 2024. 5, 7, 1
- [37] Kushal Kafle, Brian Price, Scott Cohen, and Christopher Kanan. Dvqa: Understanding data visualizations via question answering. In *CVPR*, pages 5648–5656, 2018. 4
- [38] Mehran Kazemi, Hamidreza Alvari, Ankit Anand, Jialin Wu, Xi Chen, and Radu Soricut. Geomverse: A systematic evaluation of large models for geometric reasoning. *arXiv preprint arXiv:2312.12241*, 2023. 4
- [39] Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A diagram is worth a dozen images. In *ECCV*, pages 235–251, 2016. 4
- [40] Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466, 2019. 8
- [41] Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. Race: Large-scale reading comprehension dataset from examinations. *arXiv preprint arXiv:1704.04683*, 2017. 8
- [42] Xin Lai, Zhuotao Tian, Yukang Chen, Senqiao Yang, Xiang-gu Peng, and Jiaya Jia. Step-dpo: Step-wise preference optimization for long-chain reasoning of llms. *arXiv preprint arXiv:2406.18629*, 2024. 2, 3
- [43] Hugo Laurençon, Lucile Saulnier, Léo Tronchon, Stas Bekman, Amanpreet Singh, Anton Lozhkov, Thomas Wang, Siddharth Karamcheti, Alexander Rush, Douwe Kiela, et al. Obelics: An open web-scale filtered dataset of interleaved image-text documents. *NIPS*, 36, 2024. 1
- [44] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024. 1, 6
- [45] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, pages 12888–12900, 2022. 3
- [46] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*, pages 19730–19742. PMLR, 2023. 1, 3
- [47] Lei Li, Zhihui Xie, Mukai Li, Shunian Chen, Peiyi Wang, Liang Chen, Yazheng Yang, Benyou Wang, and Lingpeng Kong. Silk: Preference distillation for large visual language models. *arXiv preprint arXiv:2312.10665*, 2023. 2, 3
- [48] Qingyun Li, Zhe Chen, Weiyun Wang, Wenhai Wang, Shenglong Ye, Zhenjiang Jin, Guanzhou Chen, Yinan He, Zhangwei Gao, Erfei Cui, et al. Omnicorpus: An unified multimodal corpus of 10 billion-level images interleaved with text. *arXiv preprint arXiv:2406.08418*, 2024. 1, 3
- [49] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *EMNLP*, pages 292–305, 2023. 6
- [50] Xi Victoria Lin, Akshat Shrivastava, Liang Luo, Sriniwasan Iyer, Mike Lewis, Gargi Gosh, Luke Zettlemoyer, and Armen Aghajanyan. Moma: Efficient early-fusion pre-training with mixture of modality-aware experts. *arXiv preprint arXiv:2407.21770*, 2024. 3
- [51] Adam Dahlgren Lindström and Savitha Sam Abraham. Clevr-math: A dataset for compositional language, visual and mathematical reasoning. *arXiv preprint arXiv:2208.05358*, 2022. 4
- [52] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*, 2023. 1, 6

- [53] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *NIPS*, 36, 2023. 1, 3, 6
- [54] Tianqi Liu, Yao Zhao, Rishabh Joshi, Misha Khalman, Mohammad Saleh, Peter J Liu, and Jialu Liu. Statistical rejection sampling improves preference optimization. *arXiv preprint arXiv:2309.06657*, 2023. 3, 5, 7, 1
- [55] Yangzhou Liu, Yue Cao, Zhangwei Gao, Weiyun Wang, Zhe Chen, Wenhai Wang, Hao Tian, Lewei Lu, Xizhou Zhu, Tong Lu, et al. Mminstruct: A high-quality multi-modal instruction tuning dataset with extensive diversity. *arXiv preprint arXiv:2407.15838*, 2024. 1, 3
- [56] Zhaoyang Liu, Yanan He, Wenhai Wang, Weiyun Wang, Yi Wang, Shoufa Chen, Qinglong Zhang, Zeqiang Lai, Yang Yang, Qingyun Li, Jiahuo Yu, et al. Interngpt: Solving vision-centric tasks by interacting with chatgpt beyond language. *arXiv preprint arXiv:2305.05662*, 2023. 3
- [57] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 1
- [58] Pan Lu, Ran Gong, Shibiao Jiang, Liang Qiu, Siyuan Huang, Xiaodan Liang, and Song-Chun Zhu. Inter-gps: Interpretable geometry problem solving with formal language and symbolic reasoning. *arXiv preprint arXiv:2105.04165*, 2021. 4
- [59] Pan Lu, Liang Qiu, Jiaqi Chen, Tony Xia, Yizhou Zhao, Wei Zhang, Zhou Yu, Xiaodan Liang, and Song-Chun Zhu. Iconqa: A new benchmark for abstract diagram understanding and visual language reasoning. *arXiv preprint arXiv:2110.13214*, 2021. 4
- [60] Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafford, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. *NIPS*, 35:2507–2521, 2022. 4
- [61] 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. 1, 2, 5, 6
- [62] Gen Luo, Xue Yang, Wenhan Dou, Zhaokai Wang, Jifeng Dai, Yu Qiao, and Xizhou Zhu. Mono-intervl: Pushing the boundaries of monolithic multimodal large language models with endogenous visual pre-training. *arXiv preprint arXiv:2410.08202*, 2024. 3
- [63] Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *CVPR*, pages 3195–3204, 2019. 4
- [64] Ahmed Masry, Xuan Long Do, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. In *ACL*, pages 2263–2279, 2022. 4
- [65] Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. Docvqa: A dataset for vqa on document images. In *WACV*, pages 2200–2209, 2021. 4
- [66] Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and CV Jawahar. Infographicvqa. In *WACV*, pages 1697–1706, 2022. 4
- [67] Igor Melnyk, Youssef Mroueh, Brian Belgodere, Matia Rigotti, Apoorva Nitsure, Mikhail Yurochkin, Kristjan Greenewald, Jiri Navratil, and Jerret Ross. Distributional preference alignment of llms via optimal transport. *arXiv preprint arXiv:2406.05882*, 2024. 5, 7, 1
- [68] Anand Mishra, Shashank Shekhar, Ajeet Kumar Singh, and Anirban Chakraborty. Ocr-vqa: Visual question answering by reading text in images. In *ICDAR*, pages 947–952, 2019. 4
- [69] Eric Mitchell. A note on dpo with noisy preferences & relationship to ipo, 2023. 5, 7, 1
- [70] OpenAI. Gpt-4v(ision) system card. https://cdn.openai.com/papers/GPTV_System_Card.pdf, 2023. 1
- [71] OpenAI. Gpt-4o system card. <https://openai.com/index/gpt-4o-system-card/>, 2024. 6
- [72] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35: 27730–27744, 2022. 3
- [73] Arka Pal, Deep Karkhanis, Samuel Dooley, Manley Roberts, Siddhartha Naidu, and Colin White. Smaug: Fixing failure modes of preference optimisation with dpo-positive. *arXiv preprint arXiv:2402.13228*, 2024. 5
- [74] Richard Yuanzhe Pang, Weizhe Yuan, Kyunghyun Cho, He He, Sainbayar Sukhbaatar, and Jason Weston. Iterative reasoning preference optimization. *arXiv preprint arXiv:2404.19733*, 2024. 2, 3
- [75] Renjie Pi, Tianyang Han, Wei Xiong, Jipeng Zhang, Runtao Liu, Rui Pan, and Tong Zhang. Strengthening multimodal large language model with bootstrapped preference optimization. *arXiv preprint arXiv:2403.08730*, 2024. 2
- [76] 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, 2024. 2, 3, 5, 7, 1
- [77] Sylvestre-Alvise Rebuffi, Hakan Bilen, and Andrea Vedaldi. Learning multiple visual domains with residual adapters. *NIPS*, 30, 2017. 3
- [78] Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy Lillicrap, Jean-baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024. 1, 6
- [79] Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. Object hallucination in image captioning. *arXiv preprint arXiv:1809.02156*, 2018. 7
- [80] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *NIPS*, 35:25278–25294, 2022. 1

- [81] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017. 3
- [82] Minjoon Seo, Hannaneh Hajishirzi, Ali Farhadi, Oren Etzioni, and Clint Malcolm. Solving geometry problems: Combining text and diagram interpretation. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 1466–1476, 2015. 4
- [83] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *CVPR*, pages 8317–8326, 2019. 4
- [84] Kai Sun, Dian Yu, Dong Yu, and Claire Cardie. Investigating prior knowledge for challenging chinese machine reading comprehension. *Transactions of the Association for Computational Linguistics*, 8:141–155, 2020. 8
- [85] Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liang-Yan Gui, Yu-Xiong Wang, Yiming Yang, et al. Aligning large multimodal models with factually augmented rlhf. *arXiv preprint arXiv:2309.14525*, 2023. 2, 3, 6, 7
- [86] Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V Le, Ed H Chi, Denny Zhou, , and Jason Wei. Challenging big-bench tasks and whether chain-of-thought can solve them. *arXiv preprint arXiv:2210.09261*, 2022. 8
- [87] Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models. *arXiv preprint arXiv:2405.09818*, 2024. 3
- [88] Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023. 1
- [89] InternLM Team. Internlm: A multilingual language model with progressively enhanced capabilities. <https://github.com/InternLM/InternLM>, 2023. 1
- [90] Bart Thomee, David A Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. Yfcc100m: The new data in multimedia research. *Communications of the ACM*, 59(2):64–73, 2016. 1
- [91] Changyao Tian, Xizhou Zhu, Yuwen Xiong, Weiyun Wang, Zhe Chen, Wenhai Wang, Yuntao Chen, Lewei Lu, Tong Lu, Jie Zhou, et al. Mm-interleaved: Interleaved image-text generative modeling via multi-modal feature synchronizer. *arXiv preprint arXiv:2401.10208*, 2024. 3
- [92] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 1
- [93] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023. 1
- [94] Binghai Wang, Rui Zheng, Lu Chen, Yan Liu, Shihan Dou, Caishuang Huang, Wei Shen, Senjie Jin, Enyu Zhou, Chenyu Shi, et al. Secrets of rlhf in large language models part ii: Reward modeling. *arXiv preprint arXiv:2401.06080*, 2024. 3
- [95] Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Mingjie Zhan, and Hongsheng Li. Measuring multimodal mathematical reasoning with math-vision dataset. *arXiv preprint arXiv:2402.14804*, 2024. 5, 6
- [96] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model’s perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 1, 3, 6
- [97] Weihaan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, et al. Cogvlm: Visual expert for pretrained language models. *arXiv preprint arXiv:2311.03079*, 2023. 3
- [98] Weiyun Wang, Yiming Ren, Haowen Luo, Tiantong Li, Chenxiang Yan, Zhe Chen, Wenhai Wang, Qingyun Li, Lewei Lu, Xizhou Zhu, et al. The all-seeing project v2: Towards general relation comprehension of the open world. *arXiv preprint arXiv:2402.19474*, 2024. 1, 3, 6
- [99] Weiyun Wang, Min Shi, Qingyun Li, Wenhai Wang, Zhenhang Huang, Linjie Xing, Zhe Chen, Hao Li, Xizhou Zhu, Zhiguo Cao, et al. The all-seeing project: Towards panoptic visual recognition and understanding of the open world. In *ICLR*, 2024. 1, 3
- [100] Weiyun Wang, Shuibo Zhang, Yiming Ren, Yuchen Duan, Tiantong Li, Shuo Liu, Mengkang Hu, Zhe Chen, Kaipeng Zhang, Lewei Lu, et al. Needle in a multimodal haystack. *arXiv preprint arXiv:2406.07230*, 2024. 3
- [101] Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Zhen Li, Qiyang Yu, et al. Emu3: Next-token prediction is all you need. *arXiv preprint arXiv:2409.18869*, 2024. 3
- [102] Yue Wu, Zhiqing Sun, Huizhuo Yuan, Kaixuan Ji, Yiming Yang, and Quanquan Gu. Self-play preference optimization for language model alignment. *arXiv preprint arXiv:2405.00675*, 2024. 5, 7, 1
- [103] Zhiheng Xi, Wenxiang Chen, Boyang Hong, Senjie Jin, Rui Zheng, Wei He, Yiwen Ding, Shichun Liu, Xin Guo, Junzhe Wang, et al. Training large language models for reasoning through reverse curriculum reinforcement learning. *arXiv preprint arXiv:2402.05808*, 2024. 2
- [104] Zhen Yang, Jinhao Chen, Zhengxiao Du, Wenmeng Yu, Weihaan Wang, Wenyi Hong, Zhihuan Jiang, Bin Xu, Yuxiao Dong, and Jie Tang. Mathglm-vision: Solving mathematical problems with multi-modal large language model. *arXiv preprint arXiv:2409.13729*, 2024. 3
- [105] Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, et al. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*, 2024. 1, 6
- [106] 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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13807–13816, 2024. 2, 3
- [107] Tianyu Yu, Haoye Zhang, Yuan Yao, Yunkai Dang, Da Chen, Xiaoman Lu, Ganqu Cui, Taiwen He, Zhiyuan Liu, Tat-Seng Chua, et al. Rlaif-v: Aligning mllms through open-source ai feedback for super gpt-4v trustworthiness. *arXiv preprint arXiv:2405.17220*, 2024. 2, 3, 6, 7
- [108] Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*, 2023. 6
- [109] Renrui Zhang, Xinyu Wei, Dongzhi Jiang, Yichi Zhang, Ziyu Guo, Chengzhuo Tong, Jiaming Liu, Aojun Zhou, Bin Wei, Shanghang Zhang, et al. Mavis: Mathematical visual instruction tuning. *arXiv preprint arXiv:2407.08739*, 2024. 3
- [110] Xiaotian Zhang, Chunyang Li, Yi Zong, Zhengyu Ying, Liang He, and Xipeng Qiu. Evaluating the performance of large language models on gaokao benchmark. *arXiv preprint arXiv:2305.12474*, 2023. 8
- [111] Yongting Zhang, Lu Chen, Guodong Zheng, Yifeng Gao, Rui Zheng, Jinlan Fu, Zhenfei Yin, Senjie Jin, Yu Qiao, Xu-anjing Huang, et al. Spa-vl: A comprehensive safety preference alignment dataset for vision language model. *arXiv preprint arXiv:2406.12030*, 2024. 2, 3
- [112] Rui Zheng, Shihan Dou, Songyang Gao, Yuan Hua, Wei Shen, Binghai Wang, Yan Liu, Senjie Jin, Qin Liu, Yuhao Zhou, et al. Secrets of rlhf in large language models part i: Ppo. *arXiv preprint arXiv:2307.04964*, 2023. 3
- [113] Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. Instruction-following evaluation for large language models. *arXiv preprint arXiv:2311.07911*, 2023. 8
- [114] Wanrong Zhu, Jack Hessel, Anas Awadalla, Samir Yitzhak Gadre, Jesse Dodge, Alex Fang, Youngjae Yu, Ludwig Schmidt, William Yang Wang, and Yejin Choi. Multimodal c4: An open, billion-scale corpus of images interleaved with text. *NIPS*, 36, 2024. 1

Enhancing the Reasoning Ability of Multimodal Large Language Models via Mixed Preference Optimization

Supplementary Material

7. Implementation Details

During the construction of samples with clear ground truths, we sample at most 32 reasoning processes and construct at most 15 preference pairs for each query. When constructing data using DropoutNTP, we truncate the original response by half and ask InternVL2-8B to complete the response without the image input. Our ablation studies in Section 8.2 show that truncating the original response by 25% or 75% has negative effects on the final performance. We set the temperature to 1.0 during sampling to ensure response diversity. Besides, the maximum tiles for dynamic resolution are set to 6 for the general VQA domain and 12 for OCR-, document-, and chart-related domains.

During the MPO process, the global batch size is set to 256 during training. We employ the AdamW optimizer [57] with the β_1 of 0.9, the β_2 of 0.999, and the weight decay of 0.05. The learning rate is initialized as $5e-6$. The training phases include a linear warmup that lasts until the first 5% of training steps. The warmup is followed by a cosine decay strategy with a minimum learning rate of 0. The KL penalty coefficient β is set to 0.1. For the Equation 3, we set w_p to 0.8, w_q to 0.2, and w_g to 1. The model is initialized from InternVL2-8B [20], and all parameters are trainable during training. We train the model for 1 epoch.

8. More Ablation Studies

8.1. Ablation Studies about DPO variants

In this section, we present the numerical experimental results of ablation studies on the effects of different preference optimization algorithms in Table 6 and Table 7. We define Δ as the performance gap between CoT reasoning responses and direct-answer responses to quantitatively assess the effects of different preference optimisation algorithms on CoT reasoning abilities. Our results indicate that introducing an additional SFT loss can significantly improve the CoT performance compared to each algorithm’s vanilla counterpart. Note that, to reduce computational costs, we only extend the DPO variants, which exhibit superior performance in Table 6 compared to DPO, with SFT Loss.

In addition to the ablation studies based on M3CoT, we also present the performance of models trained with DPO+ and BCO+ using our MMPr, as shown in Table 8. The experimental results show that models trained with MPO exhibits superior overall performance compared to those trained with DPO+ and BCO+.

Method	Direct	CoT	Δ
InternVL2-8B	59.3	57.0	-2.3
SFT	65.7	68.5	+2.8
DPO [76]	75.8	72.7	-3.1
RSO [54]	74.2	74.3	+0.1
IPO [4]	72.8	73.1	+0.3
cDPO [69]	76.2	76.8	+0.6
RobustDPO [21]	75.1	74.2	-0.9
BCO [36]	78.1	78.4	+0.3
SPPO [102]	66.2	67.4	+1.2
AOT [67]	76.7	76.0	-0.7
TR-DPO [28]	75.9	66.8	-9.1

Table 6. Results of models trained with different preference optimization algorithms on M3CoT. Δ represents the performance gap between CoT responses and direct-answer responses.

Method	Direct	CoT	Δ
ORPO [32]	66.6	73.9	+7.3
DPO+	76.4	78.9	+2.5
cDPO+	71.6	74.2	+2.7
RobustDPO+	76.5	78.0	+1.5
BCO+	77.4	78.4	+1.0
AOT+	76.3	78.0	+1.7
MPO	77.7	79.1	+1.4

Table 7. Results of models trained with different preference optimization algorithms extended with SFT Loss on M3CoT. The algorithm X extended with the SFT Loss is called X+ for brevity. For instance, DPO+ is the combination of DPO and SFT loss.

8.2. Ablation Studies on DropoutNTP

Here, we present the ablation results for the Dropout Ratio (DR) in our proposed DropoutNTP. By default, we set DR to 0.5, which means that we truncate the positive response by half. Notably, setting DR to 0.25 means using the first quarter of the positive responses for continuation. Following the experimental settings in Section 5.2.2, we replace the negative samples in RLAIIF-V with the completions based on different dropout ratios. As shown in Table 9, the model trained with data generated using a DR of 0.75 performs the worst. We attribute this to the fact that, with the first three-quarters of the prefix being identical, the difference in quality between the chosen and rejected responses becomes less apparent, reducing training effectiveness. Additionally, the model trained with a DR of 0.25

Model Name	Reasoning			General VQA		Hallucination Evaluation		
	M3CoT	MathVista	MathVision	MMVet	LLaVA-Bench	POPE	CRPE	MMHalBench
InternVL2-8B	59.3	58.3	20.4	54.2	73.2	86.9	75.0	3.3
InternVL2-8B-DPO+	80.4	66.4	23.4	58.3	74.1	87.6	75.5	3.4
InternVL2-8B-BCO+	79.6	66.1	18.8	55.5	78.6	88.5	75.5	3.5
InternVL2-8B-MPO	79.2	67.0	25.7	56.2	76.7	88.1	75.4	3.5

Table 8. Results of models trained with DPO+, BCO+ and MPO using our MMR.

Method	Object HalBench		MM HalBench	
	Resp. (\downarrow)	Ment. (\downarrow)	Score	Hall. (\downarrow)
DR=0.25	9.3	4.8	3.3	40.6
DR=0.50	7.6	4.1	3.6	31.3
DR=0.75	11.6	6.2	3.3	36.5

Table 9. Results of DropNTP with different Dropout Ratios.

performs worse than that trained with a dropout ratio of 0.5. We believe this is because the majority of the content in the rejected responses is generated without image input, resulting in noticeably lower quality compared to the chosen responses, which similarly hampers the training effectiveness. Therefore, we set the DR to 0.5.

8.3. Effects of data scale.

To evaluate the effects of the data scale, we train the model with different amounts of preference reasoning data sampled from M3CoT [16]. The M3CoT training set contains 7,861 samples annotated with corresponding rationales. To control the data volume, we adjust the maximum number of preference pairs generated for each sample, resulting in datasets of different sizes: 10K, 40K, 70K, and 100K. As illustrated in Figure 4a, model accuracy consistently improves with the increasing data volume. As the data volume rises to 100K, the model achieves its highest accuracy of 76.4 when directly answering the final answer and 78.9 when answering with CoT. Furthermore, both the Direct and CoT performance exhibit a positive correlation between data scale and accuracy, with the CoT performance achieving higher performance across all scales. These results highlight the importance of scaling up reasoning preference data to improve model performance.

8.4. Effects of hyper-parameters.

We conduct ablation studies on M3CoT to study the impact of the hyper-parameters, including learning rate, PO coefficient w_p , w_q , and SFT coefficient w_g . For the PO coefficient, we control the sum of w_p and w_q to equal 1.0 and adjust different proportions. Unless specifically mentioned, we set the learning rate to $5e-6$, w_p to 0.8, w_q to 0.2, and w_g to 1. As shown in Figure 4b, the learning rate significantly affects the model’s performance. With a rel-

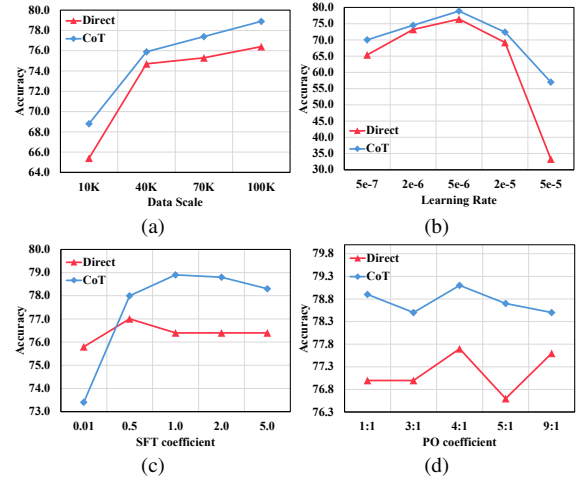


Figure 4. Results of models trained with different data scales or hyper-parameters on M3CoT. The X-axis represents the corresponding data scale or hyper-parameter for this point, while the Y-axis indicates the accuracy on M3CoT.

atively low learning rate of $5e-7$, the model shows moderate improvement. As the learning rate increases to $5e-6$, the model’s performance improves further, reaching optimal results across the tested learning rates and surpassing the baseline by 19.6 points. However, further increasing the learning rate to $5e-5$ causes a drastic performance drop, suggesting that a higher learning rate may lead to overfitting or instability in training. Additionally, the PO coefficient w_0 , w_1 and SFT coefficient w_2 are crucial. As shown in Figure 4c and 4d, the model achieves optimal performance with w_p set to 0.8, w_q set to 0.2, and w_g set to 1. Notably, when w_g is set to 0.01, the performance of the CoT approach is inferior to that of directly answering the final answer, indicating the importance of the SFT Loss during the direct preference optimization.

9. More Data Examples in MMR

In this section, we provide data examples in MMR for each task described in Table 1. Specifically, Figure 5a to 5f are examples from data constructed using DropoutNTP, while Figure 5g to 5j are examples from data constructed using correctness-based pipeline.

General VQA



Question: What are the toppings on the pizza?

Chosen Response:

The pizza in the image has the following toppings:

1. **Tomatoes:** There are halved cherry tomatoes scattered across the pizza, adding a fresh and juicy element.
 2. **Artichokes:** Pieces of artichoke hearts are visible, providing a tender and slightly tangy flavor.
 3. **Cheese:** A layer of melted cheese covers the pizza, likely mozzarella or a similar type, providing a creamy and rich texture.
 4. **Pesto or Herb Sauce:** There appears to be a green sauce, possibly pesto, spread beneath the toppings, adding a herby and savory flavor.
- The pizza has a rustic appearance with a slightly charred crust, suggesting it was baked at a high temperature.

Rejected Response:

The pizza in the image has the following toppings:

1. **Tomatoes:** There are halved cherry tomatoes scattered across the pizza, adding a fresh and juicy element.
 2. **Artichokes:** Pieces of artichoke hearts are visible, providing a tender and slightly tangy flavor.
 3. **Cheese:** A layer of melted cheese covers the pizza, likely mozzarella or a mozzarella blend, providing richness and an anchor for the other toppings.
 4. **Peppers:** Slices of green bell pepper contribute both color and crunch.
 5. **Olives:** Kalamata olives are scattered across the pizza, adding a briny and slightly salty taste.
 6. **Onion:** Thinly sliced onions contribute a sweet and pungent flavor.
- These toppings create a delicious balance of flavors and textures, making the pizza a crowd-pleaser!

(a)

Science

Question:

Two magnets are placed as shown. Hint: Magnets that attract pull together. Magnets that repel push apart. Will these magnets attract or repel each other?

A. attract B. repel



Chosen Response:

In the image, there are two magnets. Each magnet has two poles labeled as "N" for North and "S" for South.

Magnetic Interaction:

- Like Poles Repel: If similar poles (North-North or South-South) face each other, they repel.
- Unlike Poles Attract: If opposite poles (North-South) face each other, they attract.

Analysis of the Image:

1. Left Magnet: The North pole is on the left, and the South pole is on the right.
2. Right Magnet: The North pole is on the left, and the South pole is on the right.

Orientation:

- The North pole of the left magnet is facing the North pole of the right magnet.
- The South pole of the left magnet is facing the South pole of the right magnet.

Conclusion: Since both pairs of like poles (North-North and South-South) are facing each other, the magnets will repel each other.

Therefore, the correct answer is: B. repel

Rejected Response:

In the image, there are two magnets. Each magnet has two poles labeled as "N" for North and "S" for South.

Magnetic Interaction:

- Like Poles Repel: If similar poles (North-North or South-South) face each other, they repel.
- Unlike Poles Attract: If opposite poles (North-South) face each other, they attract.

Analysis of the Image:

1. Left Magnet: The North pole is on the left, and the South pole is on the right.
2. Right Magnet: The North pole is on the right, and the South pole is on the left.

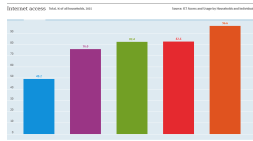
These arrangements show that the North poles are on opposite ends of the magnets, and so are the South poles.

Conclusion: Since a North pole on one magnet faces a South pole on the other, and a South pole on one magnet faces a North pole on the other: The magnets will repel each other because the North poles repel and the South poles repel.

Answer: B. repel

(b)

Chart



Question: How many color does the graph shows?

Chosen Response:

The graph uses five distinct colors, each representing a different country. Here's a detailed breakdown:

1. Blue - Represents Brazil, with an internet access percentage of 49.2%.
 2. Purple - Represents Latvia, with an internet access percentage of 76%.
 3. Green - Represents Austria, with an internet access percentage of 59.9%.
 4. Red - Represents Mexico, with an internet access percentage of 41.5%.
 5. Orange - Represents Norway, with an internet access percentage of 96.6%.
- Each color is used to differentiate the countries and their corresponding internet access percentages for 2015.

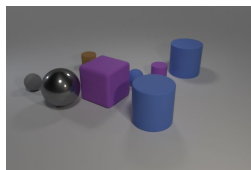
Rejected Response:

The graph uses five distinct colors, each representing a different country. Here's a detailed breakdown:

1. Blue - Represents Brazil, with an internet access percentage of 49.2%.
 2. Purple - Represents Latvia, with an internet access percentage of 76%.
 3. Green - Represents Austria, with an internet access percentage of 59.9%.
 4. Orange - Represents Mexico, with an internet access percentage of 41.5%.
 5. Red - Represents New Zealand, with an internet access percentage of 85%.
- Each country is distinctly colored and plotted at different points on the graph, which allows us to differentiate between the various countries and their corresponding internet access percentages.

(c)

Mathematics



Question: Subtract all balls. How many objects are left?

Chosen Response:

To determine how many objects are left after subtracting all the balls, let's examine the image and identify each object:

1. Balls:
 - There are three spherical objects (balls):
 - A small grey ball on the far left.
 - A shiny metallic ball next to the small grey ball.
 - A medium-sized blue ball near the center.
 2. Other Objects:
 - A purple cube, which is a square-shaped object.
 - A large blue cylinder in the foreground.
 - A small orange cylinder behind the purple cube.
 - A small purple cylinder next to the large blue cylinder.
 - A medium-sized blue cylinder in the background.
- Total Objects: - Total objects initially present = 8 (3 balls + 5 non-ball objects).
- Subtracting the Balls: - Removing the 3 balls leaves us with the non-ball objects.
- Remaining Objects: - 1 purple cube - 3 cylinders (1 large blue, 1 small orange, 1 small purple, and 1 medium blue).
- Therefore, after subtracting all the balls, there are 5 objects left.

Rejected Response:

To determine how many objects are left after subtracting all the balls, let's examine the image and identify each object:

1. Balls:
 - There are three spherical objects (balls):
 - A small grey ball on the far left.
 - A shiny metallic ball next to the small grey ball.
 - A medium-sized blue ball near the center.
 2. Other Objects:
 - A purple cube, which is a square-shaped object.
 - A large blue cylinder in the foreground.
 - A small orange cylinder behind the purple cube.
 - A small purple cylinder next to the large blue cylinder.
 - A medium-sized blue cylinder in the background.
- Step-by-step Calculation:
1. Count the total number of objects: There are five distinct objects in total.
 2. Subtract the number of balls: There are three balls in total.
 3. Calculate the remaining objects: - Total objects: 5 - Subtract the number of balls: 3 - Remaining objects: 5 - 3 = 2
- Thus, after subtracting all the balls, there are two objects left.

(d)

OCR



Question: What is the name of the event?

Chosen Response:

The event is part of the "RBC Summer Kids Stuff." It includes two programs:

1. **Camp Kilimanjaro:** This is a Vacation Bible School (VBS) program for children in Kindergarten through 5th grade. It is themed as "An Epic Expedition Through Proverbs."
 2. **Cross Training:** This is for junior high students in 6th through 8th grade, organized by Rancho Baptist Church.
- The event is scheduled from June 22 to June 26, with activities running from 8:45 AM to 12:00 PM. Registration is open, and the cost is free.

Rejected Response:

The event is part of the "RBC Summer Kids Stuff." It includes two programs:

1. **Camp Kilimanjaro:** This is a Vacation Bible School (VBS) program for children in Kindergarten through 5th grade. It is themed as "An Epic Expedition Through Proverbs."
2. **Cross Training:** This is a middle school and high school student summer program with multiple activities, including sports, recreation, community service, and Bible lessons.

(e)

Document

Question: What is the name of entity mentioned in the given form ?

Chosen Response:

The form is a "Principal Investigator's Statement of Economic Interests" related to research funding. It details information about a research project led by Paul Saltman from the Biology department at the University of California, San Diego.

The specific section of interest is Section A, which requires details about the non-governmental entity providing funding for the research. According to the form:

- Name of Entity: **Proctor & Gamble**
 - Address of Entity: **Cincinnati, Ohio**
 - Principal Business of Entity: **Household products and foods**
 - Amount of Funding: **\$25,000.00 (Actual)**
- This section is crucial as it discloses the financial backing from a non-governmental source, which is essential for transparency and compliance in research funding.

Rejected Response:

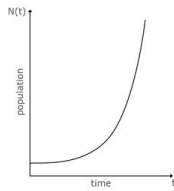
The form is a "Principal Investigator's Statement of Economic Interests" related to research funding. It details information about a research project led by Paul Saltman from the Biology department at the University of California, San Diego.

The specific section of interest is Section A, which requires details about the non-governmental entity providing funding for the project. The information in this section pertains to the funding organization, their role, and the type of funds provided. The organization mentioned is the "Cystic Fibrosis Foundation® (CFF®)". So, the name of the entity mentioned in the form is the Cystic Fibrosis Foundation® (CFF®).

The CFF® provides grants to support innovative and promising research projects in the field of cystic fibrosis, a genetic disorder affecting the lungs and digestive system. The funding primarily supports research aimed at improving patient outcomes, disease understanding, and treatments. In this case, Dr. Saltman's project aims to explore the molecular mechanisms underlying bone development during the fetus stages, potentially connecting them to cystic fibrosis and other conditions.

(f)

Science



Question:

Your task is to answer the question below. Give step by step reasoning before you answer, and when you're ready to answer, please use the format "Final answer: .."

Question:
Read the text. A population's growth is affected by factors in the environment, such as space, available food, predators, and disease. When a population inhabits an environment with abundant resources and few limiting factors, it can experience exponential growth. Under these conditions, a population grows increasingly rapidly. Plotted on a graph with time on the x-axis and population size on the y-axis, exponential growth resembles a J-shaped curve. Logistic growth, in contrast, occurs when resources are scarce or a population faces considerable limiting factors, such as predators. Logistic growth resembles an S-shaped curve: it rises steeply at first but then levels off. When growth levels off, the population has reached the environment's carrying capacity, or the population limit it can support.

Which term matches the picture?

- A. logistic growth
- B. exponential growth

Chosen Response:

To determine which term matches the picture, let's analyze the image and the descriptions provided:

1. Exponential Growth: - Description: When a population inhabits an environment with abundant resources and few limiting factors, it can experience exponential growth. - Graph Shape: Exponential growth resembles a J-shaped curve.

2. Logistic Growth: - Description: Logistic growth occurs when resources are scarce or a population faces considerable limiting factors, such as predators. - Graph Shape: Logistic growth resembles an S-shaped curve: it rises steeply at first but then levels off.

Image Analysis: - The graph shows a J-shaped curve where the population grows exponentially with time, without any leveling off.

Based on the description and the shape of the graph, the term that matches the picture is:

Final answer: B

Rejected Response:

To determine which term matches the picture, let's analyze the characteristics of each type of growth:

1. Exponential Growth: This type of growth occurs when a population inhabits an environment with abundant resources and few limiting factors. In this case, the population grows increasingly rapidly, leading to a J-shaped curve. The growth is rapid and unbounded over time.

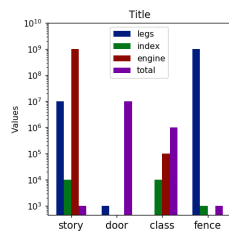
2. Logistic Growth: This type of growth occurs when resources are scarce or a population faces considerable limiting factors, such as predators. In this scenario, the population rises steeply at first but then levels off as it reaches the carrying capacity, which is the environment's maximum supportable population level. The growth curve resembles an S-shaped curve: it increases rapidly, reaches a limit, and then stabilizes.

The graph in the picture shows a population that grows rapidly at first and then levels off, resembling an S-shaped curve. This is characteristic of logistic growth.

Final answer: A

(g)

Chart



Question:

Your task is to answer the question below. Give step by step reasoning before you answer, and when you're ready to answer, please use the format "Final answer: .."

Question: What is the value of total in fence?

Chosen Response:

To determine the value of the "total" category in the "fence" row, we need to look at the color-coded legend and the corresponding values for each category in the "fence" row.

1. The "fence" row has the following categories: - legs: blue - index: green - engine: red - total: purple
2. According to the legend, the value for "total" is represented by the purple color.
3. By examining the "fence" row, we can see that the purple square is in the "total" column.
4. The purple square is aligned with the value on the right side of the image, which corresponds to 10^3 .

Therefore, the value of the "total" in the "fence" row is 10^3 .

Final answer: 1000

Rejected Response:

To determine the value of the total in fence, we need to analyze the data given for each category in the "fence" row:

1. legs: The value is 10^9 (one billion).
2. index: The value is 10^3 (one thousand).
3. engine: The value is 10^5 (one hundred thousand).

4. total: The value is not directly given, but we can infer it by summing the values of legs, index, and engine.

Let's calculate the total: total=legs+index+engine, $total = 10^9 + 10^3 + 10^5$, total = $10^9 + 10^3 + 10^5$, total = $10^9 + 10^3 + 10^5$, ... (Repeat until truncated.)

(h)

OCR

Question:

Your task is to answer the question below. Give step by step reasoning before you answer, and when you're ready to answer, please use the format "Final answer: .."

Question: What is the total amount of this receipt?

Chosen Response:

To find the total amount on the receipt, let's follow the steps provided:

1. Identify the items listed:
 - Chicken Katsu Bento (BEN02): 1, RM 12.90
 - Omu Curry Fried Rice Set (Set10): 1, RM 6.90
 - Katsu Don (Fish) Set (TS02 Kat): 1, RM 9.90
 2. Calculate the subtotal:

$$12.90 + 6.90 + 9.90 = 29.70$$
 3. Check for any discounts, service charges, or rounding adjustments:
 - There is a "Sub Total Before Discount" of RM 29.70.
 - There is no mention of any service charge or service discount.
 - The "Rounding Adjustment" is RM 0.00.
 4. Calculate the total amount:
 - The total sales amount is already provided, which is RM 29.70.
- Thus, the total amount of this receipt is RM 29.70.

Final answer: RM 29.70

Rejected Response:

To find the total amount on the receipt, let's break down the expenses listed on the receipt.

1. Chicken Katsu Bento (Beno2):
 - Quantity (QTY): 1 - Unit Price (RM): 12.90 - Discount (%): 0% - Amount (RM): 12.90
 2. Omu Curry Fried Rice Set (Set1):
 - Quantity (QTY): 1 - Unit Price (RM): 6.90 - Discount (%): 0% - Amount (RM): 6.90
 3. Katsu Don (Fish) Set (Ts02 Kats):
 - Quantity (QTY): 1 - Unit Price (RM): 9.90 - Discount (%): 0% - Amount (RM): 9.90
- Now, let's add up these amounts:
- Chicken Katsu Bento: RM 12.90
 - Omu Curry Fried Rice Set: RM 6.90
 - Katsu Don Set: RM 9.90

Total before any discounts or service charges: $12.90 + 6.90 + 9.90 = 29.70$

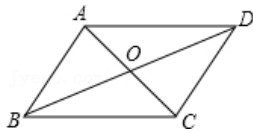
The receipt also mentions a "Sub Total Before Discount" of RM 29.70, which matches our calculation. Looking further, there are other details indicating adjustments: - Service Charge: RM 0.00 - Total Sales: RM 29.70 - Rounding Adjustment: RM 0.00

Finally, the receipt shows cash payment was RM 30.00 and the change was RM 0.30.

Final answer: 30.00

(i)

Mathematics



Question:

Your task is to answer the question below. Give step by step reasoning before you answer, and when you're ready to answer, please use the format "Final answer: ..."

Question:

Let quadrilateral ABCD have intersecting diagonals AC and BD at point O, and satisfy the conditions $AC+BD=16$ and $CD=6$. What expression represents the perimeter of triangle ABO? Use variables to represent the sides and diagonals to derive the expression.

Choices: A: 8 B: 10 C: 12 D: 14

Chosen Response:

Step-by-Step Solution:

1. Identify the Given Information:

- Quadrilateral ABCD has diagonals AC and BD intersecting at point O.
- The sum of the lengths of the diagonals is $AC + BD = 16$.
- The length of side CD is $CD = 6$.

2. Properties of the Diagonals:

- Since O is the intersection point of the diagonals, we can denote the segments of the diagonals as follows:

$AO = CO$, $BO = DO$.

3. Express the Diagonals in Terms of Segments: Let $AO = CO = x$ and $BO = DO = y$. Therefore, $AC = 2x$ and $BD = 2y$.

4. Use the Given Sum of Diagonals:

- From the given information, $AC + BD = 16$.

- Substituting the expressions for AC and BD, we get: $2x + 2y = 16$

- Simplifying, we find: $x + y = 8$

5. Perimeter of Triangle ABO : The perimeter of triangle ABO is the sum of the lengths of its sides: Perimeter of $\triangle ABO = AO + BO + AB$. We know $AO = x$ and $BO = y$, so: Perimeter of $\triangle ABO = x + y + AB$

6. Determine the Length of AB : Since $AB = CD$ (opposite sides of a parallelogram are equal), and $CD = 6$, we have: $AB = 6$

7. Calculate the Perimeter: Substitute $x + y = 8$ and $AB = 6$ into the perimeter formula:

Perimeter of $\triangle ABO = 8 + 6 = 14$

Final Answer: D

Rejected Response:

Let's break down the problem step by step:

1. Identify the given sides of quadrilateral ABCD: AC (one diagonal) + BD (the other diagonal) = 16 units. CD (one side of quadrilateral ABCD) = 6 units.

2. Understand the problem context: The diagonals AC and BD intersect at point O. We need to find the perimeter of triangle ABO.

3. Identify missing lengths: The length of AB, AD, and BO are required to find the perimeter of triangle ABO. There is no direct information about these lengths, so we need to find relationships or additional lengths that can help us.

4. Use the given information to set up an equation: Since the diagonals of a cyclic quadrilateral intersect at a point that divides each diagonal into segments that are proportional, we can use the concept of similar triangles or segments' properties.

5. Apply the intersecting chords theorem (if quadrilateral ABCD were cyclic and AC and BD its diagonals): In a cyclic quadrilateral, the product of the segments of one chord is equal to the product of the segments of the other chord. Let $AB = x$, $AD = y$, $BO = z$. As the diagonals intersect at O, triangle ADO and ABO will be similar by AA similarity.

6. Calculate the perimeter of triangle ABO: Perimeter of triangle ABO = $AB + AO + BO = x + y/2 + z$.

7. Since the full lengths of AB and BO are unknown, we need to use the lengths provided: From the given $CD = 6$ units, we can use triangles or properties to find lengths. Since no further lengths are provided, assume they are proportional using segments or similarity properties.

8. Conclusion: Given $AC + BD = 16$ units and knowing that we don't need the specific lengths of segments due to the given perimeter problem context: The triangles can potentially be found using segment lengths from given proportionalities. Given all these relationships, knowing the lengths provided, and without further calculation: The correct option is 14, Final answer: D

(j)

Figure 5. **More data examples from MMPR.** Figure 5a to 5f are examples from data constructed using DropoutNTP, while Figure 5g to 5j are examples from data constructed using correctness-based pipeline.