

---

# MLLM-as-a-Judge: Assessing Multimodal LLM-as-a-Judge with Vision-Language Benchmark

---

Dongping Chen<sup>\*1</sup> Ruoxi Chen<sup>\*2</sup> Shilin Zhang<sup>\*1</sup> Yaochen Wang<sup>\*1</sup> Yinuo Liu<sup>\*1</sup> Huichi Zhou<sup>\*1</sup>  
Qihui Zhang<sup>\*1</sup> Yao Wan<sup>1</sup> Pan Zhou<sup>1</sup> Lichao Sun<sup>3</sup>

## Abstract

Multimodal Large Language Models (MLLMs) have gained significant attention recently, showing remarkable potential in artificial general intelligence. However, assessing the utility of MLLMs presents considerable challenges, primarily due to the absence of multimodal benchmarks that align with human preferences. Drawing inspiration from the concept of LLM-as-a-Judge within LLMs, this paper introduces a novel benchmark, termed MLLM-as-a-Judge, to assess the ability of MLLMs in assisting judges across diverse modalities, encompassing three distinct tasks: *Scoring Evaluation*, *Pair Comparison*, and *Batch Ranking*. Our study reveals that, while MLLMs demonstrate remarkable human-like discernment in *Pair Comparison*, there is a significant divergence from human preferences in *Scoring Evaluation* and *Batch Ranking*. Furthermore, a closer examination reveals persistent challenges in the judgment capacities of LLMs, including diverse biases, hallucinatory responses, and inconsistencies in judgment, even in advanced models such as GPT-4V. These findings emphasize the pressing need for enhancements and further research efforts to be undertaken before regarding MLLMs as fully reliable evaluators. In light of this, we advocate for additional efforts dedicated to supporting the continuous development within the domain of MLLM functioning as judges. The code and dataset are publicly available at our project homepage: <https://mllm-judge.github.io/>.

---

<sup>\*</sup>Equal contribution <sup>1</sup>Huazhong University of Science and Technology <sup>2</sup>Zhejiang University of Technology <sup>3</sup>LAIR Lab, Lehigh University. Correspondence to: Yao Wan <wanyao@hust.edu.cn>, Pan Zhou <panzhou@hust.edu.cn>.

*Proceedings of the 41<sup>st</sup> International Conference on Machine Learning*, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

## 1. Introduction

The advent of Large Language Models (LLMs), such as GPT-3 (OpenAI, 2023) and Llama (Touvron et al., 2023), has achieved substantial progress in content generation, including text generation (OpenAI, 2023), code generation (Roziere et al., 2023), and video synthesis (Wu et al., 2023a). The emergent abilities of LLMs, as demonstrated by the Chain-of-Thought (CoT) framework (Wei et al., 2022), present a promising avenue for their utilization as evaluators, also referred to as the LLM-as-a-Judge (Zheng et al., 2023b). Initial explorations indicate a better alignment with human preferences, emphasizing the considerable potential inherent in this approach.

Recently, building upon LLMs, Multimodal Large Language Models (MLLMs) like GPT-4V (OpenAI, 2023) and LLaVA (Liu et al., 2023d) exhibit exceptional proficiency by incorporating multiple modalities (e.g., text, charts, images, and videos) and showcasing remarkable performance in multimodal applications, including text-to-video (Wu et al., 2023a) and visual dialog (Cai et al., 2023). Despite this, assessing the effectiveness of MLLMs remains challenging due to the limitations of traditional metrics, which hinge on text-based exact matches or embedding distances. These metrics fall short in adhering to the granular evaluation criteria of interest and fail to capture the rich context within the generated outputs. Drawing inspiration from the concept of LLM-as-a-Judge within LLMs, a pertinent research question arises: “Can MLLMs effectively serve as judges in the multimodal domain, and how closely do their evaluations align with human preferences?”

To answer this question, this paper undertakes an extensive study, introducing a groundbreaking benchmark, MLLM-as-a-Judge, specifically crafted to evaluate the efficacy of MLLMs in assisting judges across diverse modalities. To achieve this goal, we first thoughtfully curate a selection of 14 datasets across various tasks, including image captioning, math reasoning, text reading, and infographics understanding, culminating in acquiring a dataset comprising 4,414 image-instruction pairs. Subsequently, we utilize six mainstream MLLMs from a model pool which includes GPT-4V

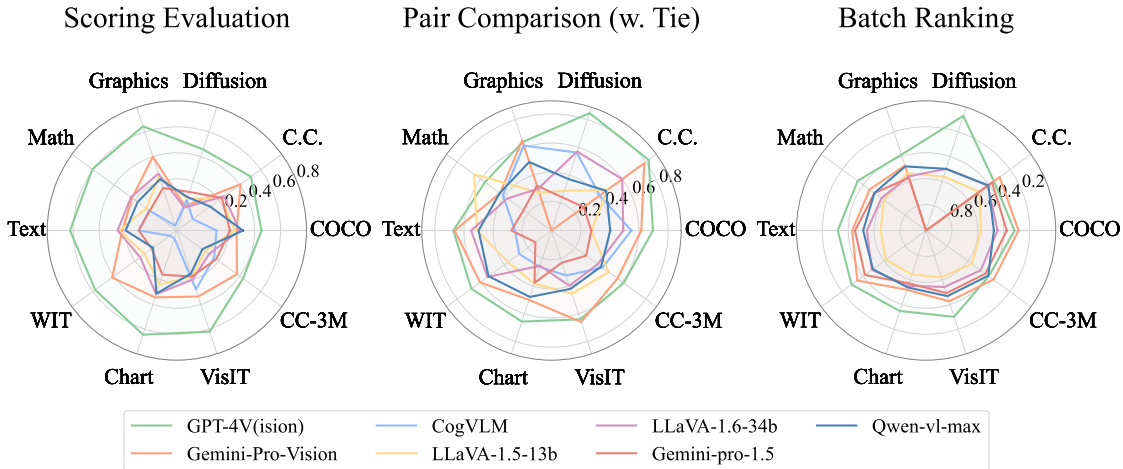


Figure 1. Comparative performance of different MLLMs across three judging settings in 10 datasets, each is the average of three iterations. As the CogVLM is unable to perform the batch ranking task, we show the other six MLLMs only.

(OpenAI, 2023), Gemini (GeminiTeam, 2023)<sup>1</sup>, LLaVA-1.5-13b, LLaVA-1.6-34b (Liu et al., 2023d), CogVLM (Wang et al., 2023c), Qwen-VL-Max (Bai et al., 2023a), to generate responses to each instruction across three distinct evaluation settings. The produced responses are subsequently gathered and undergo additional annotation by human evaluators, who apply stringent criteria to ensure an impartial and thorough assessment of the judgments made by the MLLMs.

Furthermore, we assess the ability of MLLMs as judges in multimodal tasks by calculating the similarity between human and MLLMs judgment and measuring human agreement on the analysis and judgment made by those MLLMs. In particular, we target eleven widely-used MLLMs, i.e., GPT-4V and Gemini-Pro-1.0/1.5, CogVLM, LLaVA-1.5/1.6 family, and Qwen-VL family, across two settings (with, or without vision input), over three distinct tasks (i.e., *Scoring Evaluation*, *Pair Comparison*, and *Batch Ranking*). Figure 1 compares the performance of various MLLMs across different datasets and settings, illustrating that GPT-4V exhibits significantly superior capabilities as a judge compared to other MLLMs.

As a benchmark, we also release two curated datasets to facilitate further studies: MLLM-AS-A-JUDGE-HQ, which showcases responses with a high level of concordance with human judgments, and MLLM-AS-A-JUDGE-HARD, which includes responses marked by inconsistency with human preferences and instances of hallucination. Additionally, we address the limitations of MLLMs in judgment, such as egocentric bias, position bias, length bias, and hallucination. We demonstrate that integrating CoT (Wei et al., 2022) and a vision expert system can effectively mitigate some of these biases.

<sup>1</sup>For conciseness, we refer to GPT-4V(ision) as GPT-4V, and Gemini-Pro-Vision as Gemini throughout this paper.

**Take-Aways.** We evaluate the judgment performance of 11 MLLMs across 14 datasets under three settings: score evaluation, pair comparison, and batch ranking. Our findings reveal several key insights. First, while MLLMs demonstrate proficiency in aligning with human preferences in pair comparison tasks, they require further improvement in score evaluation and batch ranking, particularly in reasoning tasks. Secondly, GPT-4V consistently outperforms other models across all tasks and settings. Finally, the presence of hallucinations, biases, and inconsistent judgments in MLLMs highlights significant challenges that must be addressed for these models to become a viable alternative to traditional human evaluations.

To summarize, our work provides three key contributions:

- **A Benchmark.** We are the first to develop a comprehensive benchmark MLLM-as-a-Judge in multimodal domains, with human annotations to assess the judging capability of MLLMs in tasks of *Scoring Evaluation*, *Pair Comparison* and *Batch Ranking*.
- **Two Datasets.** We curate two human preference datasets: MLLM-AS-A-JUDGE-HQ, which contains high-quality questions, and MLLM-AS-A-JUDGE-HARD, which includes instances of hallucination. These datasets can serve as rigorous testing grounds to facilitate the development of MLLMs in aligning human preferences.
- **Findings and Implications.** Our evaluation of mainstream MLLMs reveals that while MLLMs exhibit alignment with human judgments in *Pair Comparison*, notable discrepancies can be found in *Scoring Evaluation* and *Batch Ranking*. Furthermore, our findings reveal that MLLMs exhibit a range of biases and hallucinations, along with inconsistent judgments during the evaluation process, representing significant hurdles in establishing MLLMs as reliable judges.

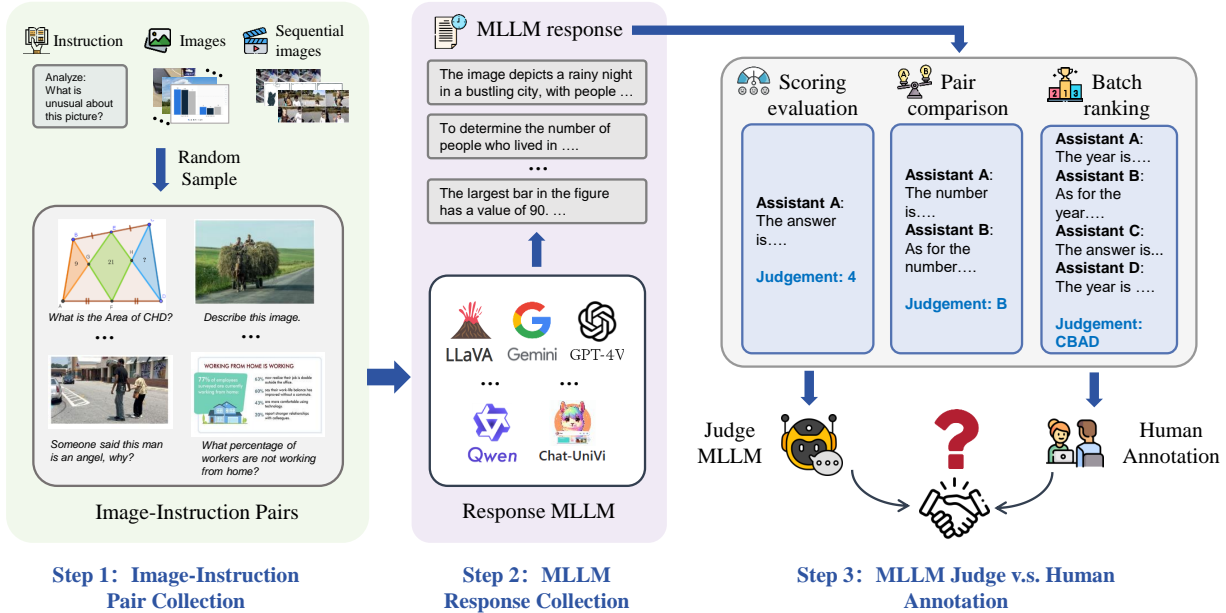


Figure 2. An overview of MLLM-as-a-Judge.

## 2. MLLM-as-a-Judge: A Benchmark to Assess Vision-Language Judging Ability

Figure 2 shows an overview of our proposed MLLM-as-a-Judge, consisting of three steps: 1) image-instruction pair collection, 2) MLLM response collection, and 3) comparison with human annotation. Initially, we collect a dataset  $\mathcal{P} = \{(M_1, I_1), \dots, (M_n, I_n)\}$ , containing pairs of images ( $M$ ) and their corresponding instructions ( $I$ ) sourced from 10 diverse domains (e.g., math, chart, diffusion), ensuring comprehensive coverage for a wide array of downstream tasks. Subsequently, each pair  $(M_i, I_i)$  is processed through several MLLMs, generating a set of responses  $\mathcal{R}_i = \{r_1, r_2, \dots, r_n\}$  for each pair. This process contributes to the formation of the dataset of image-instruction-responses pairs, denoted as  $\mathcal{D} = \{(M_i, I_i, \mathcal{R}_i) | (M_i, I_i) \in \mathcal{P}\}$ . Finally, the dataset  $\mathcal{D}$  is partitioned into three distinct subsets to facilitate diverse task evaluations:  $\mathcal{D}_{\text{score}}$  for *Scoring Evaluation*,  $\mathcal{D}_{\text{pair}}$  for *Pair Comparison*, and  $\mathcal{D}_{\text{batch}}$  for *Batch Ranking*. Each subset will be employed for specific judging tasks, with each of them being configured as follows.

- **Scoring Evaluation:** Each individual response is evaluated on a scale from 1 to 5, with the specific criteria for this rating system detailed in Appendix F.
- **Pair Comparison:** It involves a direct comparison between two responses, culminating in the identification of the superior one. Following the principles outlined by (Deutsch et al., 2023), a tie option is incorporated to ensure a more equitable assessment.
- **Batch Ranking:** The responses are systematically arranged in descending order of quality based on a given instruction, without any tie option.

### 2.1. Step 1: Image-Instruction Pair Collection

We meticulously curate a dataset consisting of 4,414 image-text pairs, gathered from a variety of downstream task datasets, as detailed in Table 8 in Appendix B. These pairs are carefully tailored into image-instruction pairs to suit a free-form response format. To illustrate, within the domain of diffusion tasks, our dataset incorporated pairs challenging models to adeptly recognize and articulate connections between provided images and user-specified keywords.

### 2.2. Step 2: MLLM Response Collection

We employ six widely-used MLLMs – GPT-4V (OpenAI, 2023), Gemini (GeminiTeam, 2023), LLaVA (Liu et al., 2023d), Qwen-VL-Max (Bai et al., 2023a), LLaVA-1.6-34b (Liu et al., 2023d), and CogVLM (Wang et al., 2023c) – to generate responses based on the image-instruction pairs, obtaining approximately 17,000 responses. Responses that are either too brief or non-compliant with security regulations (e.g., “I’m sorry, but I cannot assist with this request”) from GPT-4V and Gemini are excluded. The number of responses and the length distributions for different MLLMs are shown in Table 1 and Figure 3, respectively. We show specific hyper-parameter settings in Appendix B.2. Besides, we segment these responses into three non-overlapping groups, to prevent response overlap.

### 2.3. Step 3: Comparison with Human Annotations

The annotation is conducted by 6 authors of this paper independently. These annotators are proficient in this domain, with different genders, ages, and educational backgrounds to

Table 1. The statistics of responses in different steps for MLLM judging. In Step 3, under the w.o. vision input settings, we sample 10% from the original data and mainly proceed with GPT-4V and Gemini. We only list the amount of judgments generated by four models here. M-I: Image-Instruction.

Step	Setting	Input	Num.	Output	Num.			
1	/	Image	4,144	M-I Pairs	4,400			
		Instruction	4,414					
2	/	M-I Pairs	3,300	MLLMs	17,096			
3	w. Vision Input	Batch	1,470	Gemini	1,340			
				GPT-4V	1,454			
				Qwen-VL-Max	1,458			
				LLaVA	1,468			
		Pair	8,256	Gemini	7,751			
				GPT-4V	8,117			
	Qwen-VL-Max			8,012				
	LLaVA			8,253				
	Score	5,883	Gemini	5,337				
			GPT-4V	5,708				
	w.o. Vision Input	No Vision	Batch	110	Gemini	107		
					GPT-4V	110		
Pair					425	Gemini	385	
Vision Experts		Pair	425	425	GPT-4V	355		
					Score	612	Gemini	582
					Score	612	612	GPT-4V
Batch	110	Gemini	107					
		GPT-4V	110					
Pair	425	425	Gemini	396				
			GPT-4V	425				
Score	612	612	Gemini	576				
			GPT-4V	612				

ensure diversity (Sun et al., 2020). They are required to give objective judgments without considering answer lengths, and certain names or positions of the response to minimize human bias. More details are referred to Appendix E.

### 3. Experiment Settings

#### 3.1. Settings of MLLM-as-a-Judge

We evaluate the judging performance of eleven leading MLLMs – GPT-4V (OpenAI, 2023), Gemini-Pro-Vision-1.0 (GeminiTeam, 2023), LLaVA-1.5-13b, LLaVA-1.6-7b/13b/34b (Liu et al., 2023d), Qwen-VL-Plus/Max (Bai et al., 2023a) and CogVLM (Wang et al., 2023c) – across three distinct evaluation settings. Adapting the “Analyze-then-Judge” paradigm from Chiang & Lee (2023b), which is a one-step CoT approach (Wei et al., 2022), we first ask MLLMs to analyze responses and then provide a judgment based on their analysis. However, due to capability limitations to perform the “Analyze-then-Judge” setting for LLaVA and CogVLM, we prompt them to directly output their judgment. We also evaluate whether multi-step CoT

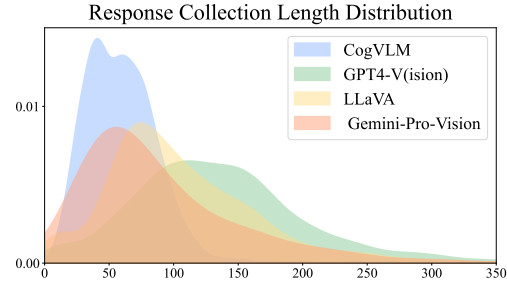


Figure 3. Length distribution in responses for different MLLMs. Horizontal axis: length; Vertical axis: density.

will enhance the performance of MLLM serving as a judge.

Furthermore, to extensively explore MLLMs judging capabilities, we conduct experiments on various settings, including scenarios without vision input, replacing vision input with a detailed description generated by GPT-4V as a vision expert, and employing multi-step CoT. Considering that the first two settings do not involve image inputs, we also include tests on the latest GPT-4 (OpenAI, 2023) Gemini (GeminiTeam, 2023), LLaMA-2-70b (Touvron et al., 2023), and Mixtral-8x7b (Jiang et al., 2024) to assess whether LLMs can effectively perform judging tasks without vision perception. Comprehensive details of these experimental setups are available in Appendix C, and the prompts can be found in Appendix F.

#### 3.2. Judging Metrics

After collecting responses from MLLM judgments, we quantify their alignment with human annotations across three settings, employing distinct metrics as follows:

▷ **Scoring Evaluation:** Following LLM-as-a-Judge (Zheng et al., 2023b), we compute the Pearson similarity (Lee Rodgers & Nicewander, 1988) between the MLLMs’ judgments and human ratings across different sub-datasets.

▷ **Pair Comparison:** We assess the similarity between the MLLM judgments and human decisions using accuracy, F1-score (Goutte & Gaussier, 2005), and recall (Goutte & Gaussier, 2005) to assess the judging abilities of models.

▷ **Batch Evaluation:** We consolidate the ranking results into a singular sequence and employ the Normalized Levenshtein distance (Levenshtein et al., 1966) to evaluate the similarity between judgments from MLLMs and human annotation.

#### 3.3. Human Agreement in MLLM Judgment

Apart from traditional metrics for similarity assessment between judgments from MLLMs and humans, we further evaluate the judgments provided by MLLMs to uncover latent bias and hallucination in 10 datasets. We also invite human annotators for further validation, focusing on the following aspects:

Table 2. The overall performance of different MLLMs in judging, compared with human annotations on different datasets. We sample all the data three times and took the average to mitigate the casualty. *w.* and *w.o.* tie represents tie and non-tie situations respectively. We omit Gemini’s results on the diffusion task for its challenges in processing AI-generated images. All presented data of Pearson similarity exhibit a  $p$ -value below 0.05, indicating a statistically significant level of confidence. Please refer to the Appendix D.1 for more results.

Settings	MLLM	COCO	C.C.	Diff.	Graphics	Math	Text	WIT	Chart	VisIT	CC-3M	M2W	SciQA	Aes	MM-Vet	Ave.
Score (†)	LLaVA-1.5-13b	0.247	0.227	0.060	0.242	0.093	0.245	0.109	0.237	0.177	0.071	<b>0.424</b>	0.279	<b>0.414</b>	0.322	0.225
	LLaVA-1.6-34b	0.285	0.251	-0.012	0.262	0.238	0.258	0.151	0.318	0.198	0.109	0.022	0.206	0.025	0.265	0.184
	Gemini	0.262	0.408	-	0.400	0.228	0.222	0.418	0.343	0.336	0.374	0.324	0.073	0.360	0.207	0.304
	GPT-4V	<b>0.454</b>	<b>0.507</b>	<b>0.458</b>	<b>0.645</b>	<b>0.606</b>	<b>0.624</b>	<b>0.579</b>	<b>0.645</b>	<b>0.620</b>	<b>0.431</b>	0.185	<b>0.383</b>	0.401	<b>0.326</b>	<b>0.490</b>
	Qwen-vl-max	0.311	0.117	0.072	0.218	0.175	0.196	0.028	0.312	0.151	0.045	0.244	0.115	0.177	0.216	0.170
Pair w. Tie (†)	LLaVA-1.5-13b	0.273	0.478	0.286	0.273	<b>0.657</b>	0.510	0.369	0.383	0.456	0.484	0.347	0.223	0.389	0.254	0.384
	LLaVA-1.6-34b	0.493	0.600	0.570	0.300	0.374	0.551	0.543	0.254	0.398	0.392	0.513	<b>0.434</b>	0.524	0.499	0.460
	Gemini	0.616	0.787	-	<b>0.650</b>	0.436	0.664	0.605	0.500	<b>0.660</b>	0.560	0.370	0.262	0.190	0.312	0.509
	GPT-4V	<b>0.696</b>	<b>0.824</b>	<b>0.847</b>	0.639	0.564	<b>0.673</b>	<b>0.679</b>	<b>0.657</b>	0.640	<b>0.612</b>	<b>0.521</b>	0.415	<b>0.606</b>	<b>0.529</b>	<b>0.636</b>
	Qwen-vl-max	0.403	0.464	0.372	0.494	0.438	0.500	0.533	0.479	0.421	0.421	0.411	0.392	0.325	0.474	0.438
Pair w.o. Tie (†)	LLaVA-1.5-13b	0.327	0.537	0.302	0.300	0.726	0.684	0.600	0.610	0.648	0.583	0.449	0.443	0.498	0.344	0.504
	LLaVA-1.6-34b	0.607	0.824	0.855	0.402	0.587	0.750	<b>0.758</b>	0.381	0.503	0.564	<b>0.712</b>	<b>0.679</b>	0.694	<b>0.762</b>	0.648
	Gemini	0.717	0.840	-	0.770	0.678	0.793	0.688	0.658	0.711	0.652	0.471	0.358	0.265	0.400	0.615
	GPT-4V	<b>0.804</b>	<b>0.870</b>	<b>0.922</b>	<b>0.807</b>	<b>0.801</b>	<b>0.805</b>	<b>0.734</b>	<b>0.849</b>	<b>0.761</b>	<b>0.703</b>	0.699	0.647	<b>0.755</b>	0.659	<b>0.773</b>
	Qwen-vl-max	0.657	0.674	0.556	0.667	0.635	0.732	0.647	0.638	0.560	0.586	0.608	0.646	0.741	0.662	0.644
Batch (↓)	LLaVA-1.5-13b	0.577	0.492	0.562	0.535	0.598	0.650	0.616	0.644	0.620	0.563	0.639	0.563	0.650	0.652	0.597
	LLaVA-1.6-34b	0.449	0.411	0.500	0.561	0.575	0.544	0.483	0.552	0.542	0.479	<b>0.529</b>	0.437	0.500	0.450	0.501
	Gemini	<b>0.287</b>	<b>0.299</b>	-	0.473	0.462	0.430	0.344	0.520	0.426	0.357	0.613	<b>0.412</b>	0.467	0.529	0.432
	GPT-4V	0.318	0.353	<b>0.070</b>	<b>0.385</b>	<b>0.348</b>	<b>0.319</b>	<b>0.290</b>	<b>0.347</b>	<b>0.300</b>	<b>0.402</b>	0.597	0.462	0.453	<b>0.411</b>	<b>0.361</b>
	Qwen-vl-max	0.477	0.407	0.500	0.480	0.507	0.515	0.493	0.539	0.468	0.407	0.563	0.503	<b>0.444</b>	0.500	0.486

▷ **Human Agreement:** This involves a simple ‘yes’ or ‘no’ response to assess agreement with the MLLM judgments. While some judgments might appear reasonable, they may still be considered incorrect due to unique human perspectives. Hence, we conduct experiments on human agreement to address situations that traditional metrics may not adequately capture.

▷ **Analysis Grading:** Each MLLM analysis is assigned a score from 1 to 5, considering relevance, accuracy, creativity, and response granularity, detailed in Appendix F.

▷ **Hallucination Detection:** Given the propensity for hallucination issues in the complex reasoning chains and long-term vision-language contexts of MLLMs, we task human annotators with identifying any hallucinations in the analyses of MLLM judgments, adhering to established definitions of vision and language hallucination (Sun et al., 2024).

## 4. Empirical Results and Analysis

### 4.1. MLLM Judgment vs Human Annotation

As shown in Figure 1 and Table 3, judgments made by GPT-4V are closer to human annotations among all settings, while Gemini is far different, with LLaVA, CogVLM and Qwen-VL-Max are even worse. Overall, MLLM judgments perform better on *Pair Comparison*, while falling short in *Scoring Evaluation* and *Batch Ranking*, showing a huge gap between the model and human preferences. Under the ‘Analyze-then-Judge’ setting, GPT-4V prefers to give a

longer judge in all settings, convincing its ability to reason on long-term text.

▷ **Scoring Evaluation:** GPT-4V demonstrates the highest similarity to human scoring with a similarity score of 0.490. In contrast, Gemini achieves only 0.304, with LLaVA and CogVLM scoring even lower. This discrepancy is mainly due to Gemini’s tendency to assign scores around 4 points as depicted in Figure 4, seldom giving 1 or 2 points. LLaVA and CogVLM show a pattern similar to Gemini, predominantly assigning scores around 4 points. We attribute this to a ‘High-Score’ Bias, akin to the ‘Yes/No’ bias identified by Liu et al. (2023a), which may result from an imbalance in positive and negative judging instructions in their training data (Liu et al., 2023b), severely limits their ability to provide just and varied scores in scoring settings. In comparison, GPT-4V’s scores are more evenly distributed and align closely with human preferences.

▷ **Pair Comparison:** As illustrated in Figure 4, GPT-4V outshines other MLLMs in pair comparison tasks, achieving 0.636 in tie settings and 0.773 in non-tie settings, surpassing 0.8 in many datasets, which indicate a strong alignment with human preferences. Gemini, LLaVA, and CogVLM show a marked preference for declaring a clear winner, possibly due to a lack of tie situations in their training, leading to biased judgments. It’s also interesting that the frequency of ties given by GPT-4V closely mirrors that of human judges, suggesting similar thresholds for tie decisions.

▷ **Batch Ranking:** GPT-4V aligns more closely with human

Table 3. Human agreement percentage on MLLM-as-a-Judge in 10 datasets. Each judgment is independently reviewed three times by different annotators and consensus results are recorded. Gemini failed in diffusion tasks and its results are omitted.

Settings	MLLM	COCO	C.C.	Diffusion	Graphics	Math	Text	WIT	Chart	VisIT	CC-3M	Average
Score (↑)	Gemini	0.783	<b>0.739</b>	-	0.618	0.536	0.621	<b>0.749</b>	0.630	0.712	0.702	0.677
	GPT-4V	<b>0.799</b>	0.725	<b>0.506</b>	<b>0.688</b>	<b>0.638</b>	<b>0.706</b>	0.714	<b>0.676</b>	<b>0.779</b>	<b>0.754</b>	<b>0.699</b>
Pair (↑)	Gemini	0.705	0.833	-	0.733	0.520	0.717	<b>0.827</b>	0.620	<b>0.853</b>	0.703	0.724
	GPT-4V	<b>0.821</b>	<b>0.926</b>	<b>0.873</b>	<b>0.794</b>	<b>0.618</b>	<b>0.752</b>	0.790	<b>0.796</b>	0.797	<b>0.766</b>	<b>0.793</b>
Batch (↓)	Gemini	0.642	<b>0.639</b>	-	0.333	0.330	0.473	0.511	0.315	0.422	<b>0.554</b>	0.469
	GPT-4V	<b>0.663</b>	<b>0.639</b>	<b>0.912</b>	<b>0.536</b>	<b>0.475</b>	<b>0.615</b>	<b>0.641</b>	<b>0.640</b>	<b>0.622</b>	0.467	<b>0.621</b>

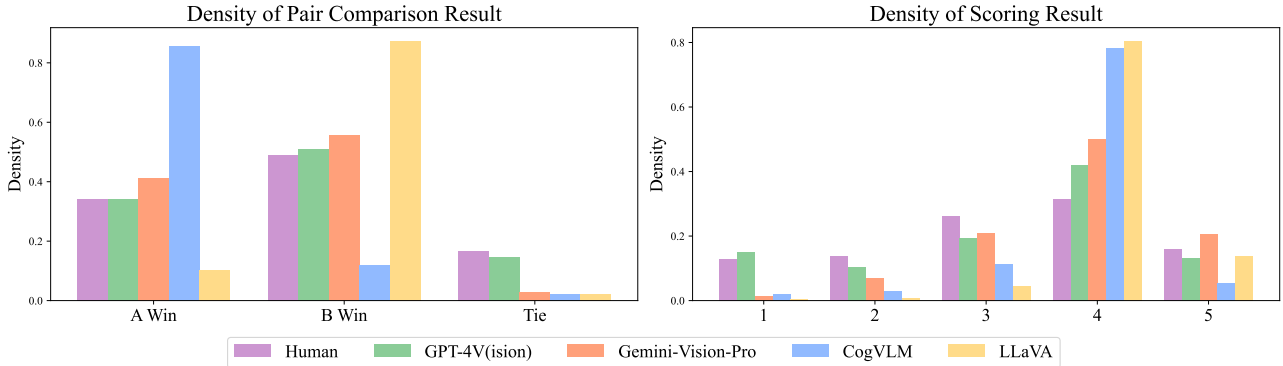


Figure 4. Pair Comparison density (Left) and Scoring Evaluation density (Right) of different MLLMs judgments and human annotations.

Table 4. Consistency comparisons of GPT-4V and Gemini in 10 datasets. Average means weighted average for consistency times, ‘‘MCC’’ stands for ‘‘Majority Consistency Criterion’’, which deems responses consistent if over half of them are identical across our 6 repetitions of experiments.

MLLM	Score		Pair		Batch	
	Average	MCC	Average	MCC	Average	MCC
Gemini	0.531	0.054	0.781	0.547	0.629	0.338
GPT-4V	<b>0.796</b>	<b>0.611</b>	<b>0.836</b>	<b>0.675</b>	<b>0.679</b>	<b>0.418</b>

ranking results, indicating a significant lead with a mean Levenshtein Distance of 0.361. However, there is still substantial room for improvement in this task for all MLLMs. Notably, CogVLM is unable to provide a full ranking in this context, offering only the top choice; so it was excluded from this comparison; LLaVA also exhibits position bias influenced by prompt structure, often replicating judgments seen in example prompts, which complicates its ability to produce fair judgments.

4.2. MLLM Judging Consistency

To be a reliable judge, consistent decision-making across repeated evaluations of the same query is crucial. For this purpose, we conduct six repeated tests with MLLM judgments and calculated the weighted average consistency scores and Majority Consistency Criterion ratios for GPT-4V and

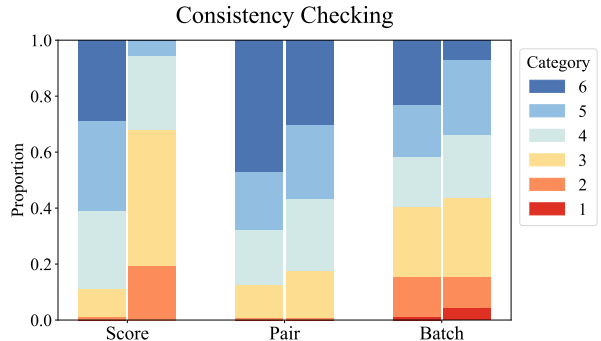


Figure 5. Consistency checking on 6 repetitions of experiments on GPT-4V (Left) and Gemini (Right). GPT-4V outperforms Gemini with a relatively higher ratio for high consistency.

Gemini, as shown in Table 4 and Figure 5. Despite a higher temperature setting, GPT-4V substantially outperforms Gemini across all tasks. Particularly in Pair Comparison, GPT-4V achieves a higher consistency score of 0.675, but it encounters difficulties in maintaining similar levels of consistency in Scoring and Batch Ranking tasks, with scores dropping to 0.611 and 0.418, indicating the challenge of producing qualified and convincing judgments.

4.3. Human Agreement

Our manual evaluation of MLLMs on agreement and scoring, revealed notable findings. Table 3 shows that GPT-

Table 5. Results of GPT-4V and Gemini-Pro acting as a judge with a 3-step CoT approach in a selected subset.

Settings	MLLM	COCO	C.C.	Diffusion	Graphics	Math	Text	WIT	Chart	VisIT	CC-3M	Ave.
Score (↑)	GPT-4V	<b>0.454</b>	<b>0.507</b>	<b>0.458</b>	<b>0.645</b>	<b>0.606</b>	<b>0.624</b>	<b>0.579</b>	<b>0.645</b>	<b>0.620</b>	<b>0.431</b>	<b>0.557</b>
	GPT-4V (+CoT)	0.246	0.165	0.192	0.385	0.397	0.400	0.298	0.443	0.423	0.038	0.299
	Gemini	0.262	0.408	-	0.400	0.228	0.222	0.418	0.343	0.336	0.374	0.299
	Gemini (+CoT)	0.127	0.068	0.117	0.220	0.132	0.182	0.105	0.140	0.222	0.128	0.144
Pair w. Tie (↑)	GPT-4V	<b>0.696</b>	<b>0.824</b>	<b>0.847</b>	<b>0.639</b>	<b>0.564</b>	<b>0.673</b>	<b>0.679</b>	<b>0.657</b>	0.640	<b>0.612</b>	<b>0.683</b>
	GPT-4V (+CoT)	0.507	0.657	0.561	0.601	0.515	0.580	0.489	0.521	<b>0.646</b>	0.553	0.563
	Gemini	0.616	0.787	-	0.650	0.436	0.664	0.605	0.500	0.660	0.560	0.609
	Gemini (+CoT)	0.233	0.239	0.420	0.207	0.284	0.329	0.352	0.357	0.247	0.239	0.291
Pair w.o. Tie (↑)	GPT-4V	<b>0.804</b>	<b>0.870</b>	<b>0.922</b>	<b>0.807</b>	<b>0.801</b>	<b>0.805</b>	<b>0.734</b>	<b>0.849</b>	<b>0.761</b>	<b>0.703</b>	<b>0.806</b>
	GPT-4V (+CoT)	0.673	0.821	0.845	0.707	0.738	0.787	0.548	0.756	0.753	0.654	0.728
	Gemini	0.717	0.840	-	0.770	0.678	0.793	0.688	0.658	0.711	0.652	0.723
	Gemini (+CoT)	0.267	0.275	0.573	0.264	0.414	0.424	0.427	0.511	0.299	0.319	0.377
Batch (↓)	GPT-4V	0.323	0.344	<b>0.092</b>	<b>0.401</b>	<b>0.367</b>	<b>0.341</b>	<b>0.302</b>	<b>0.364</b>	<b>0.313</b>	0.407	<b>0.325</b>
	GPT-4V (+CoT)	0.428	0.416	-	0.427	0.434	0.401	0.366	0.406	0.422	0.472	0.419
	Gemini	<b>0.287</b>	<b>0.299</b>	-	0.473	0.462	0.430	0.344	0.520	0.426	<b>0.357</b>	0.400
	Gemini (+CoT)	0.441	0.481	0.542	0.595	0.494	0.533	0.483	0.569	0.486	0.463	0.509

4V achieved around 70% human agreement across all settings, excelling in the *Pair Comparison* task with 79.3% agreement. Specifically, GPT-4V reached 78% in human agreement for *Pair Comparison*, with Gemini close at 72%, indicating strong performance in most sample pairs and supporting the idea that large models excel in pairwise distinctions (Zheng et al., 2023b), though improvements are needed in other judging settings.

In *Scoring Evaluation*, GPT-4V achieves a 70% human agreement rate, peaking at 79.9% in MS-COCO, while Gemini averaged 67.7%. To assess the consistency of MLLM judging quality across multiple responses to a single image-instruction pair, we use Mean Absolute Deviation (MAD) metric to measure the average absolute variance between individual scores and the mean. Figure 18 shows that GPT-4V exhibits lower variation in quality assessments, indicating more consistent and reliable judgment compared to Gemini. However, in *Batch Ranking*, both models exhibited decreased alignment with human judgments, especially in Maths and graphic information processing, suggesting that models may lack the capabilities to fully comprehend user instructions, leading to less reliable judgments.

#### 4.4. Multi-steps CoT Do Not Enhance Performance

We have conducted additional tests using GPT-4V and Gemini with a 3-step CoT approach for judging, as detailed in Table 5. Our analysis reveals that while employing CoT with additional steps markedly reduces hallucinations in judgments, it does not align more closely with human preferences. On numerous datasets, this approach even diminishes judging performance. Specifically, Gemini’s effectiveness drops more drastically. With 3-step CoT, there is an increased likelihood that the judgment will be disturbed by its understanding of the figure and its own responses to

the instruction, thereby undermining its final judgment if hallucinations exist in the previous chain.

#### 4.5. Vision Perception Benefits MLLM Judging

We explore the feasibility of using LLMs for judging text-based responses without directly analyzing the original images. This involves two approaches: omitting vision information entirely and providing a detailed description of the picture. We choose LLaMA-70b, Mixtral8x7b-v0.1 and GPT-3.5 to provide descriptions. Surprisingly, as illustrated in Table 6, we find that LLMs’ performance in multimodal judging tasks significantly improve with picture descriptions, achieving a Pearson similarity of 0.435 in *Scoring Evaluation* tasks, markedly outperformed judgments made without any vision perception. Notably, in no-tie *Pair Comparison*, MLLMs with detailed vision descriptions even exceed the standard performance of MLLMs in judging. This suggests that MLLMs may lack certain human-like judging capabilities, while LLMs can be potential judges for multimodal tasks when provided with comprehensive task-related descriptions.

#### 4.6. Bias and Hallucination

**Egocentric Bias.** Models tend to assign higher scores to their own responses while scoring others lower (Zheng et al., 2023b; Li et al., 2024). In Figures 19 and 20, GPT-4V exhibits a slight degree of Egocentricity. Conversely, Gemini maintains a uniform scoring distribution across different sources, demonstrating a more equitable approach to judgment. In contrast, GPT-4V shows self-preference, aligning its judgments with its predefined ethical guidelines. For example, GPT-4V consistently emphasizes privacy preservation, leading to higher scores for privacy-related questions based on its own metrics. Despite efforts in prompt engineer-

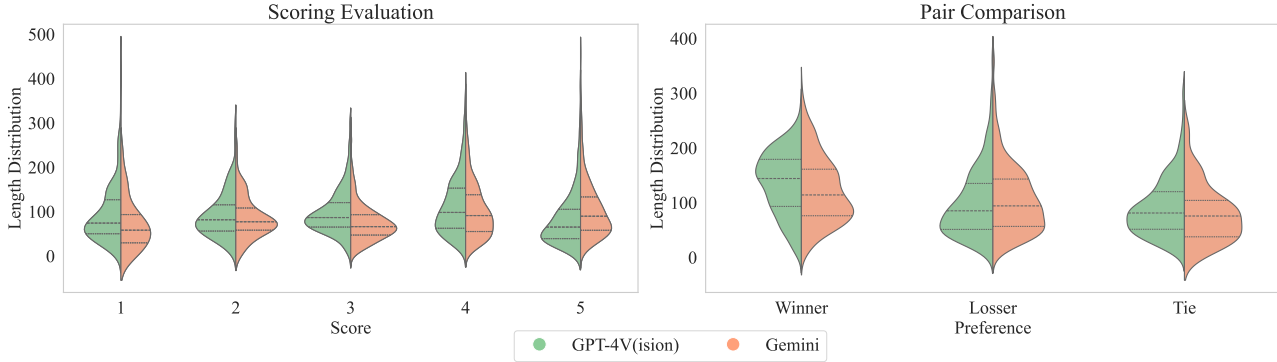


Figure 6. Length bias in 10 datasets. The horizontal axis represents length, and the vertical axis represents density.

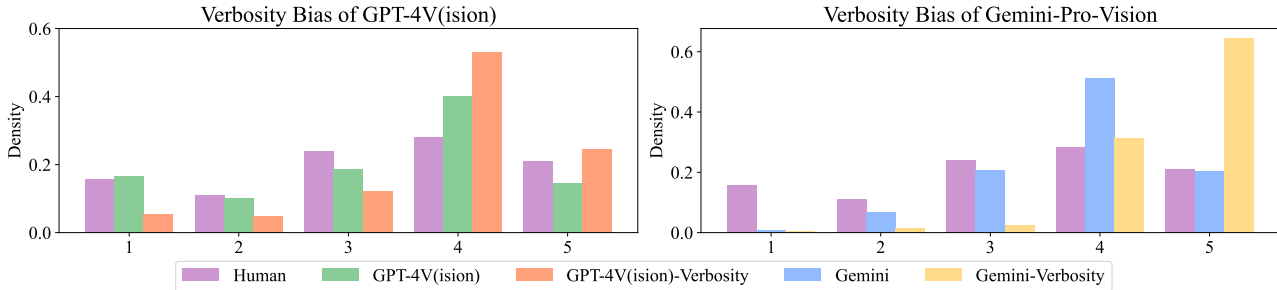


Figure 7. Length Bias in Different MLLM judgments.

Table 6. How vision perception significantly enhances multimodal judging performance in traditional LLM-as-a-Judge setting, slightly outperforming MLLMs in judging. Vision Exp. stands for judging with a detailed image description.

MLLM	Settings	Score (↑)	Pair (↑)		Batch (↓)
		Pearson	w. Tie	w.o. Tie	Edit Dis.
LLaMA2-70b	Vision Exp	0.060	0.404	0.550	0.643
	No Vision	0.126	0.374	0.537	0.583
Mixtral-8x7b	Vision Exp	0.054	0.374	0.543	0.603
	No Vision	0.151	0.478	0.731	0.546
GPT-3.5	Vision Exp	0.154	0.453	0.591	0.473
	No Vision	0.223	0.459	0.644	0.504
GPT-4V	Vision Exp	<b>0.435</b>	<b>0.544</b>	<b>0.878</b>	0.400
	No Vision	0.299	0.491	0.868	<b>0.394</b>
Gemini	Vision Exp	0.120	0.438	0.785	0.472
	No Vision	0.108	0.433	0.758	0.470

ing to ensure neutrality, these models still rely on judgment criteria set during post-alignment training (Ouyang et al., 2022). This bias can result in judgments that deviate from human preferences, highlighting the complexity of aligning MLLM judgments with humans’.

**Position Bias.** Model consistently favor answers in specific positions, often influenced by training data that typically places correct responses at the beginning or end of prompts (Liu et al., 2023e). Figure 4 illustrates bias in LLaVA and CogVLM during Pair Comparison tasks, where

they consistently prefer answers in a specific position. This bias likely arises from their limited ability to follow complex instructions, leading them to be influenced by prompt structure. For example, if a *Batch Ranking* prompt includes a sequence like ‘ABCD’, LLaVA replicates this sequence in 88.2% of responses, significantly more than other sequences. However, this bias can be reduced by introducing multiple examples, suggesting that prompts with more examples can better direct these models to follow instructions accurately.

**Length Bias.** Models tend to prefer longer answers over concise but correct ones (Li et al., 2024), also known as verbosity bias (Zheng et al., 2023b). Figure 6 shows that both GPT-4V and Gemini assign higher scores to longer content. We conducted an expanded scoring experiment using GPT-4 (OpenAI, 2023) without vision, increasing the semantic length of answers without changing their original intent. In Figure 7, we observe noticeable score increases, with GPT-4V and Gemini showing average gains of 0.6 and 0.75 points, respectively. These results suggest that MLLMs may favor longer text for higher scores.

**Hallucination Detection and Mitigation.** We observe a higher frequency of hallucinations in *Batch Ranking*, compared to *Pair Comparison* and *Scoring Evaluation*. These hallucinations involved significant misinterpretations and retrieval errors, impacting judgment accuracy and reliability. To address this, we employed a multi-step CoT approach



Table 7. Reduction of hallucinations in MLLM-AS-A-JUDGE-HARD through additional CoT steps compared to normal setting.

Setting	Figure-instruction	Figure	Instruction
Score	46.15%	<b>48.72%</b>	33.33%
Pair	28.21%	<b>35.90%</b>	33.33%
Batch	<b>43.59%</b>	35.90%	35.90%

on MLLM-AS-A-JUDGE-HARD, adding reasoning steps before the conventional “Analyze-then-Judge” process. This enhanced procedure included: 1) image-instruction, 2) image, and 3) instruction. In Table 7, this strategy effectively reduced hallucinations across all formats, with significant improvements in tasks involving image-related information. In the *Batch Ranking* task, which requires handling longer text sequences, the detailed reasoning steps were particularly effective in reducing hallucinations.

#### 4.7. Scaling Law for MLLM-as-a-Judge

We conduct two sets of experiments with models of different sizes, the LLaVA-1.6 series models and the Qwen series models in four newly added datasets, illustrated in Figure 10 and 11. In *Score evaluation*, LLaVA-1.6-34b and Qwen-VL-Max slightly outperform others in Math, Chart, and Text tasks, showing a relatively strong scaling law.

### 5. Related Work

**LLM as a Judge.** The evolution of LLMs has made them increasingly effective evaluators in Natural Language Processing (NLP) tasks. Zhu et al. (2023) introduced JudgeLM for LLM evaluation, followed by AUTO-J (Li et al., 2023a), aligning closely with human judgment (Bai et al., 2023b; Li et al., 2023d; Kim et al., 2023). Advancements in CoT reasoning (Wei et al., 2022; Chu et al., 2023) and training-free instruction following (Brown et al., 2020; Wei et al., 2021) further extend LLMs’ judging capability in diverse tasks like translation quality assessment (Kocmi & Federmann, 2023) and story generation (Chiang & Lee, 2023a).

**Hallucination and Bias in Judgments.** MLLMs suffer from vision and language hallucinations (Ji et al., 2023; Huang et al., 2023a; Cui et al., 2023; Wang et al., 2023a), often due to vision-language misalignments in training phase (Sun et al., 2024; Huang et al., 2023b). Recent research focuses on hallucination evaluation (Liu et al., 2023a), detection (Li et al., 2023e; Wang et al., 2023a), and mitigation (Yin et al., 2023; Gunjal et al., 2023; Zhou et al., 2023), noting that even GPT-4V suffer from these issues (Shi et al., 2023; Liu et al., 2023a; Cui et al., 2023). Besides, biases in MLLM-as-a-Judge, similar to those in human decision-making (Blunch, 1984; Raghbir & Valen-

zuela, 2006) and other ML domains (Wang et al., 2018; Liu et al., 2023e), such as position (Zheng et al., 2023a), egocentric (Li et al., 2024), and verbosity biases (Saito et al., 2023), are compounded by the integration of visual perception, necessitating further investigation.

### 6. Future Directions

**Multimodal RLHF/DPO.** Our work is highly connected with multimodal RLHF/DPO (Sun et al., 2023; Li et al., 2023c; Yu et al., 2023a). Our dataset includes extensive human annotations, such as manually assigned scores and preference on pairs, which could serve as invaluable training material for RLHF reward models and supply paired data essential for DPO (Rafailov et al., 2024; Zhang et al., 2024), paving the way for enhancing the training of MLLMs.

**Exploring the upper bound of MLLM-as-a-Judge.** Beyond expanding the steps in the Chain of Thought prompting (Wei et al., 2022), we see significant potential in more sophisticated reasoning frameworks, such as multi-agent debating (Chan et al., 2023) when MLLM acts as a Judge, which could enhance the judging accuracy through improved reasoning capabilities. Additionally, addressing inherent biases in the model during the judgment process is crucial. For instance, position bias in *Pair Comparison* and *Batch Ranking* (Zheng et al., 2023a; Wang et al., 2024a), and the tendency to assign higher scores, as discussed in (Lee et al., 2024), are critical areas for improvement.

Incorporating a human-in-the-loop approach (Wang et al., 2023b) offers a promising solution to enhance judgment consistency and reliability. For example, if judgment results vary in more than half of several repeated judgments, it may need human intervention for consistency checking. When it’s challenging to discern the MLLM’s judgment due to non-compliance with the suggested output format or lack of a clear outcome, human intervention may be required to refine this process by manually verifying judgments.

### 7. Conclusion

In this paper, we have presented a new benchmark, termed MLLM-as-a-Judge, to assess the judging capabilities of MLLMs across three critical evaluation settings in the multimodal domain: *Scoring Evaluation*, *Pair Comparison*, and *Batch Ranking*. We further evaluate their agreement with humans. Our results reveal that advanced MLLMs can win significant human recognition in *Pair Comparisons*, but perform poorly in *Scoring Evaluation* and *Batch Ranking* tasks. Our work highlights potential areas for future refinement and improvement of MLLMs. We advocate for additional efforts dedicated to supporting the continuous development of MLLMs as judges.

## Impact Statement

In this paper, we introduce a novel benchmark, termed MLLM-as-a-Judge, designed to propel the evolution of MLLMs toward achieving judgments that align more closely with human perspectives. This benchmark establishes a heightened criterion for assessing MLLMs, emphasizing their proficiency in comprehending and processing information in a manner reflective of human cognitive processes. One limitation of our work lies in the bias in human annotation and MLLMs. We leave the exploration of more objectives, ethically principled, and socially beneficial MLLM-as-a-Judge systems as our future work.

## References

- Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C. L., and Parikh, D. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pp. 2425–2433, 2015.
- Bai, J., Bai, S., Yang, S., Wang, S., Tan, S., Wang, P., Lin, J., Zhou, C., and Zhou, J. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 2023a.
- Bai, S., Yang, S., Bai, J., Wang, P., Zhang, X., Lin, J., Wang, X., Zhou, C., and Zhou, J. Touchstone: Evaluating vision-language models by language models. *arXiv preprint arXiv:2308.16890*, 2023b.
- Banerjee, S. and Lavie, A. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pp. 65–72, 2005.
- Bitton, Y., Bansal, H., Hessel, J., Shao, R., Zhu, W., Awadalla, A., Gardner, J., Taori, R., and Schimdt, L. Visit-bench: A benchmark for vision-language instruction following inspired by real-world use. *ArXiv*, abs/2308.06595, 2023. URL <https://api.semanticscholar.org/CorpusID:260887670>.
- Blunch, N. J. Position bias in multiple-choice questions. *Journal of Marketing Research*, 21(2):216–220, 1984.
- Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901, 2020.
- Cai, Y., Mao, S., Wu, W., Wang, Z., Liang, Y., Ge, T., Wu, C., You, W., Song, T., Xia, Y., et al. Low-code llm: Visual programming over llms. *arXiv preprint arXiv:2304.08103*, 2023.
- Chan, C.-M., Chen, W., Su, Y., Yu, J., Xue, W., Zhang, S., Fu, J., and Liu, Z. Chateval: Towards better llm-based evaluators through multi-agent debate. *arXiv preprint arXiv:2308.07201*, 2023.
- Chiang, C.-H. and Lee, H.-y. Can large language models be an alternative to human evaluations? *arXiv preprint arXiv:2305.01937*, 2023a.
- Chiang, C.-H. and Lee, H.-y. A closer look into automatic evaluation using large language models. *arXiv preprint arXiv:2310.05657*, 2023b.
- Chu, Z., Chen, J., Chen, Q., Yu, W., He, T., Wang, H., Peng, W., Liu, M., Qin, B., and Liu, T. A survey of chain of thought reasoning: Advances, frontiers and future. *arXiv preprint arXiv:2309.15402*, 2023.
- Cui, C., Zhou, Y., Yang, X., Wu, S., Zhang, L., Zou, J., and Yao, H. Holistic analysis of hallucination in gpt-4v (ision): Bias and interference challenges. *arXiv preprint arXiv:2311.03287*, 2023.
- Deng, X., Gu, Y., Zheng, B., Chen, S., Stevens, S., Wang, B., Sun, H., and Su, Y. Mind2web: Towards a generalist agent for the web. *Advances in Neural Information Processing Systems*, 36, 2024.
- Deutsch, D., Foster, G., and Freitag, M. Ties matter: Meta-evaluating modern metrics with pairwise accuracy and tie calibration. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 12914–12929, 2023.
- GeminiTeam. Gemini: A family of highly capable multimodal models, 2023.
- Goutte, C. and Gaussier, E. A probabilistic interpretation of precision, recall and f-score, with implication for evaluation. In *European conference on information retrieval*, pp. 345–359. Springer, 2005.
- Gunjal, A., Yin, J., and Bas, E. Detecting and preventing hallucinations in large vision language models. *arXiv preprint arXiv:2308.06394*, 2023.
- Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., et al. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *arXiv preprint arXiv:2311.05232*, 2023a.
- Huang, S., Hu, J., Yang, Z., Yang, L., Luo, T., Chen, H., Sun, L., and Yang, B. Decision mamba: Reinforcement learning via hybrid selective sequence modeling, 2024a.

- Huang, Y., Zhang, Q., Sun, L., et al. Trustgpt: A benchmark for trustworthy and responsible large language models. *arXiv preprint arXiv:2306.11507*, 2023b.
- Huang, Y., Yuan, Q., Sheng, X., Yang, Z., Wu, H., Chen, P., Yang, Y., Li, L., and Lin, W. Aesbench: An expert benchmark for multimodal large language models on image aesthetics perception. *arXiv preprint arXiv:2401.08276*, 2024b.
- Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., and Fung, P. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38, 2023.
- Jiang, A. Q., Sablayrolles, A., Roux, A., Mensch, A., Savary, B., Bamford, C., Chaplot, D. S., de las Casas, D., Hanna, E. B., Bressand, F., Lengyel, G., Bour, G., Lample, G., Lavaud, L. R., Saulnier, L., Lachaux, M.-A., Stock, P., Subramanian, S., Yang, S., Antoniak, S., Scao, T. L., Gervet, T., Lavril, T., Wang, T., Lacroix, T., and Sayed, W. E. Mixtral of experts, 2024.
- Jin, P., Takanobu, R., Zhang, C., Cao, X., and Yuan, L. Chat-univi: Unified visual representation empowers large language models with image and video understanding. *arXiv preprint arXiv:2311.08046*, 2023.
- Kim, S., Shin, J., Cho, Y., Jang, J., Longpre, S., Lee, H., Yun, S., Shin, S., Kim, S., Thorne, J., et al. Prometheus: Inducing fine-grained evaluation capability in language models. *arXiv preprint arXiv:2310.08491*, 2023.
- Kocmi, T. and Federmann, C. Large language models are state-of-the-art evaluators of translation quality. *arXiv preprint arXiv:2302.14520*, 2023.
- Lee, S., Kim, S., Park, S. H., Kim, G., and Seo, M. Prometheus-vision: Vision-language model as a judge for fine-grained evaluation. *arXiv preprint arXiv:2401.06591*, 2024.
- Lee Rodgers, J. and Nicewander, W. A. Thirteen ways to look at the correlation coefficient. *The American Statistician*, 42(1):59–66, 1988.
- Levenshtein, V. I. et al. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, volume 10, pp. 707–710. Soviet Union, 1966.
- Li, J., Sun, S., Yuan, W., Fan, R.-Z., Zhao, H., and Liu, P. Generative judge for evaluating alignment. *arXiv preprint arXiv:2310.05470*, 2023a.
- Li, K., Wang, Y., He, Y., Li, Y., Wang, Y., Liu, Y., Wang, Z., Xu, J., Chen, G., Luo, P., et al. Mvbench: A comprehensive multi-modal video understanding benchmark. *arXiv preprint arXiv:2311.17005*, 2023b.
- Li, L., Xie, Z., Li, M., Chen, S., Wang, P., Chen, L., Yang, Y., Wang, B., and Kong, L. Silkie: Preference distillation for large visual language models. *arXiv preprint arXiv:2312.10665*, 2023c.
- Li, X., Zhang, T., Dubois, Y., Taori, R., Gulrajani, I., Guestrin, C., Liang, P., and Hashimoto, T. B. AlpacaEval: An automatic evaluator of instruction-following models. *GitHub repository*, 2023d.
- Li, Y., Du, Y., Zhou, K., Wang, J., Zhao, W. X., and Wen, J.-R. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*, 2023e.
- Li, Z., Xu, X., Shen, T., Xu, C., Gu, J.-C., and Tao, C. Leveraging large language models for nlg evaluation: A survey. *arXiv preprint arXiv:2401.07103*, 2024.
- Lin, C.-Y. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pp. 74–81, 2004.
- Lin, T.-Y., Maire, M., Belongie, S. J., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, 2014. URL <https://api.semanticscholar.org/CorpusID:14113767>.
- Liu, F., Guan, T., Li, Z., Chen, L., Yacoob, Y., Manocha, D., and Zhou, T. HallusionBench: You see what you think? or you think what you see? an image-context reasoning benchmark challenging for gpt-4v (ision), llava-1.5, and other multi-modality models. *arXiv preprint arXiv:2310.14566*, 2023a.
- Liu, F., Lin, K., Li, L., Wang, J., Yacoob, Y., and Wang, L. Aligning large multi-modal model with robust instruction tuning. *arXiv preprint arXiv:2306.14565*, 2023b.
- Liu, H., Li, C., Li, Y., and Lee, Y. J. Improved baselines with visual instruction tuning, 2023c.
- Liu, H., Li, C., Wu, Q., and Lee, Y. J. Visual instruction tuning, 2023d.
- Liu, N. F., Lin, K., Hewitt, J., Paranjape, A., Bevilacqua, M., Petroni, F., and Liang, P. Lost in the middle: How language models use long contexts. *arXiv preprint arXiv:2307.03172*, 2023e.
- Lu, P., Mishra, S., Xia, T., Qiu, L., Chang, K.-W., Zhu, S.-C., Tafjord, O., Clark, P., and Kalyan, A. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, 35:2507–2521, 2022.

- Lu, P., Bansal, H., Xia, T., Liu, J., Yue Li, C., Hajishirzi, H., Cheng, H., Chang, K.-W., Galley, M., and Gao, J. Mathvista: Evaluating math reasoning in visual contexts with gpt-4v, bard, and other large multimodal models. *ArXiv*, abs/2310.02255, 2023. URL <https://api.semanticscholar.org/CorpusID:264491155>.
- Masry, A., Long, D., Tan, J. Q., Joty, S., and Hoque, E. ChartQA: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 2263–2279, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-acl.177. URL <https://aclanthology.org/2022.findings-acl.177>.
- Mathew, M., Bagal, V., Tito, R. P., Karatzas, D., Valveny, E., and Jawahar, C. Infographicvqa. *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 2582–2591, 2021. URL <https://api.semanticscholar.org/CorpusID:233394125>.
- OpenAI. Gpt-4 technical report. 2023.
- OpenAI. Openai models - gpt-4-vision. <https://openai.com/research/gpt-4v-system-card>, 2023.
- Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022.
- Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pp. 311–318, 2002.
- Prendki, J. Are you spending too much money labeling data?, 2023.
- Rafailov, R., Sharma, A., Mitchell, E., Manning, C. D., Ermon, S., and Finn, C. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36, 2024.
- Raghubir, P. and Valenzuela, A. Center-of-inattention: Position biases in decision-making. *Organizational Behavior and Human Decision Processes*, 99(1):66–80, 2006.
- Roziere, B., Gehring, J., Gloeckle, F., Sootla, S., Gat, I., Tan, X. E., Adi, Y., Liu, J., Remez, T., Rapin, J., et al. Code llama: Open foundation models for code. *arXiv preprint arXiv:2308.12950*, 2023.
- Saito, K., Wachi, A., Wataoka, K., and Akimoto, Y. Verbosity bias in preference labeling by large language models. *arXiv preprint arXiv:2310.10076*, 2023.
- Sharma, P., Ding, N., Goodman, S., and Soricut, R. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Annual Meeting of the Association for Computational Linguistics*, 2018. URL <https://api.semanticscholar.org/CorpusID:51876975>.
- Shi, Y., Peng, D., Liao, W., Lin, Z., Chen, X., Liu, C., Zhang, Y., and Jin, L. Exploring ocr capabilities of gpt-4v (ision): A quantitative and in-depth evaluation. *arXiv preprint arXiv:2310.16809*, 2023.
- Singh, A., Natarajan, V., Shah, M., Jiang, Y., Chen, X., Batra, D., Parikh, D., and Rohrbach, M. Towards vqa models that can read. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8309–8318, 2019. URL <https://api.semanticscholar.org/CorpusID:85553602>.
- Srinivasan, K., Raman, K., Chen, J., Bendersky, M., and Najork, M. Wit: Wikipedia-based image text dataset for multimodal multilingual machine learning. *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2021. URL <https://api.semanticscholar.org/CorpusID:232092726>.
- Sun, L., Huang, Y., Wang, H., Wu, S., Zhang, Q., Gao, C., Huang, Y., Lyu, W., Zhang, Y., Li, X., et al. Trustllm: Trustworthiness in large language models. *arXiv preprint arXiv:2401.05561*, 2024.
- Sun, W., Nasraoui, O., and Shafto, P. Evolution and impact of bias in human and machine learning algorithm interaction. *Plos one*, 15(8):e0235502, 2020.
- Sun, Z., Shen, S., Cao, S., Liu, H., Li, C., Shen, Y., Gan, C., Gui, L.-Y., Wang, Y.-X., Yang, Y., et al. Aligning large multimodal models with factually augmented rlhf. *arXiv preprint arXiv:2309.14525*, 2023.
- Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.
- Vedantam, R., Lawrence Zitnick, C., and Parikh, D. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4566–4575, 2015.

- Wang, J., Zhou, Y., Xu, G., Shi, P., Zhao, C., Xu, H., Ye, Q., Yan, M., Zhang, J., Zhu, J., et al. Evaluation and analysis of hallucination in large vision-language models. *arXiv preprint arXiv:2308.15126*, 2023a.
- Wang, P., Li, L., Chen, L., Zhu, D., Lin, B., Cao, Y., Liu, Q., Liu, T., and Sui, Z. Large language models are not fair evaluators. *arXiv preprint arXiv:2305.17926*, 2023b.
- Wang, W., Lv, Q., Yu, W., Hong, W., Qi, J., Wang, Y., Ji, J., Yang, Z., Zhao, L., Song, X., Xu, J., Xu, B., Li, J., Dong, Y., Ding, M., and Tang, J. Cogvlm: Visual expert for pretrained language models, 2023c.
- Wang, X., Golbandi, N., Bendersky, M., Metzler, D., and Najork, M. Position bias estimation for unbiased learning to rank in personal search. In *Proceedings of the eleventh ACM international conference on web search and data mining*, pp. 610–618, 2018.
- Wang, X., Ma, B., Hu, C., Weber-Genzel, L., Röttger, P., Kreuter, F., Hovy, D., and Plank, B. "my answer is c": First-token probabilities do not match text answers in instruction-tuned language models. *arXiv preprint arXiv:2402.14499*, 2024a.
- Wang, X., Zhou, Y., Liu, X., Lu, H., Xu, Y., He, F., Yoon, J., Lu, T., Bertasius, G., Bansal, M., et al. Mementos: A comprehensive benchmark for multimodal large language model reasoning over image sequences. *arXiv preprint arXiv:2401.10529*, 2024b.
- Wang, Z. J., Montoya, E., Munechika, D., Yang, H., Hoover, B., and Chau, D. H. Diffusiondb: A large-scale prompt gallery dataset for text-to-image generative models. *ArXiv*, abs/2210.14896, 2022. URL <https://api.semanticscholar.org/CorpusID:253116574>.
- Wei, J., Bosma, M., Zhao, V. Y., Guu, K., Yu, A. W., Lester, B., Du, N., Dai, A. M., and Le, Q. V. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*, 2021.
- Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q. V., Zhou, D., et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35: 24824–24837, 2022.
- Wu, S., Fei, H., Qu, L., Ji, W., and Chua, T.-S. Nextgpt: Any-to-any multimodal llm. *arXiv preprint arXiv:2309.05519*, 2023a.
- Wu, Y., Wang, S., Yang, H., Zheng, T., Zhang, H., Zhao, Y., and Qin, B. An early evaluation of gpt-4v (ision). *arXiv preprint arXiv:2310.16534*, 2023b.
- Yang, Z., Li, L., Lin, K., Wang, J., Lin, C.-C., Liu, Z., and Wang, L. The dawn of Imms: Preliminary explorations with gpt-4v (ision). *arXiv preprint arXiv:2309.17421*, 9 (1):1, 2023.
- Yin, S., Fu, C., Zhao, S., Xu, T., Wang, H., Sui, D., Shen, Y., Li, K., Sun, X., and Chen, E. Woodpecker: Hallucination correction for multimodal large language models. *arXiv preprint arXiv:2310.16045*, 2023.
- Yu, T., Yao, Y., Zhang, H., He, T., Han, Y., Cui, G., Hu, J., Liu, Z., Zheng, H.-T., Sun, M., et al. Rllhf-v: Towards trustworthy mllms via behavior alignment from fine-grained correctional human feedback. *arXiv preprint arXiv:2312.00849*, 2023a.
- Yu, W., Yang, Z., Li, L., Wang, J., Lin, K., Liu, Z., Wang, X., and Wang, L. Mm-vet: Evaluating large multimodal models for integrated capabilities. *ArXiv*, abs/2308.02490, 2023b. URL <https://api.semanticscholar.org/CorpusID:260611572>.
- Zhang, R., Gui, L., Sun, Z., Feng, Y., Xu, K., Zhang, Y., Fu, D., Li, C., Hauptmann, A., Bisk, Y., et al. Direct preference optimization of video large multimodal models from language model reward. *arXiv preprint arXiv:2404.01258*, 2024.
- Zheng, C., Zhou, H., Meng, F., Zhou, J., and Huang, M. On large language models' selection bias in multi-choice questions. *arXiv preprint arXiv:2309.03882*, 2023a.
- Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E., et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *arXiv preprint arXiv:2306.05685*, 2023b.
- Zhou, Y., Cui, C., Yoon, J., Zhang, L., Deng, Z., Finn, C., Bansal, M., and Yao, H. Analyzing and mitigating object hallucination in large vision-language models. *arXiv preprint arXiv:2310.00754*, 2023.
- Zhu, L., Wang, X., and Wang, X. Judgelm: Fine-tuned large language models are scalable judges. *arXiv preprint arXiv:2310.17631*, 2023.

## A. Comprehensive Related Works

### A.1. Large Model as Judge

The rapid development of LLMs has significantly enhanced their capabilities in long-term context perception and reasoning, increasingly popularizing their use as evaluators in various Natural Language Processing (NLP) tasks. [Zhu et al. \(2023\)](#) were pioneers in this area, introducing JudgeLM, a fine-tuned LLM designed for evaluating other LLMs. Building on this, [Li et al. \(2023a\)](#) introduced AUTO-J, a system that evaluates LLMs through both pairwise comparisons and single-response assessments, demonstrating close alignment with human judgment ([Bai et al., 2023b](#); [Li et al., 2023d](#); [Kim et al., 2023](#)). Further advancements in LLMs, such as the development of Chain-of-Thought reasoning ([Wei et al., 2022](#); [Chu et al., 2023](#)), training-free instruction following ([Brown et al., 2020](#); [Wei et al., 2021](#)), and enhanced alignment with human preferences ([Ouyang et al., 2022](#)), have solidified their role in diverse tasks like translation quality assessment ([Kocmi & Federmann, 2023](#)) and story generation ([Chiang & Lee, 2023a](#)).

### A.2. Hallucination and Bias in Judge

MLLMs are known to exhibit both vision hallucination and hallucination originating from LLMs, a phenomenon typically characterized by responses containing information not present in the visual or natural language context ([Ji et al., 2023](#); [Huang et al., 2023a](#); [Cui et al., 2023](#); [Wang et al., 2023a](#)). This issue often stems from misalignments in vision-language training ([Sun et al., 2024](#); [Huang et al., 2023b](#)). Recent studies have begun to address these hallucination issues, focusing on evaluation ([Liu et al., 2023a](#)), detection ([Li et al., 2023e](#); [Wang et al., 2023a](#)), and mitigation strategies ([Yin et al., 2023](#); [Gunjal et al., 2023](#); [Zhou et al., 2023](#)). Notably, GPT-4V ([OpenAI, 2023](#)), despite being a leading model in many fields ([Yang et al., 2023](#); [Wu et al., 2023b](#)), has also demonstrated susceptibility to hallucinations ([Shi et al., 2023](#); [Liu et al., 2023a](#); [Cui et al., 2023](#)). This raises concerns about the reliability of MLLMs in evaluative roles.

In terms of bias, MLLM judging is subject to issues not exclusive to our context of evaluation but also observed in human decision-making ([Blunch, 1984](#); [Raghubir & Valenzuela, 2006](#)) and Machine Learning (ML) domains ([Wang et al., 2018](#); [Liu et al., 2023e](#); [Huang et al., 2024a](#)) such as position bias ([Zheng et al., 2023a](#)), egocentric bias ([Li et al., 2024](#)), and verbosity bias ([Saito et al., 2023](#)). The integration of visual perception in MLLMs introduces additional complexities, resulting in biases unique to the fusion of dual perceptions, an area that still demands thorough exploration.

### A.3. Evaluating Large Multimodal Models

Evaluating MLLMs typically involves diverse tasks and corresponding metrics, which reflect the models' ability to comprehend and generate content based on both visual and textual information. For instance, in image captioning tasks, models are tasked with generating descriptive text for a given image. The effectiveness of these models is measured using metrics such as BLEU ([Papineni et al., 2002](#)), METEOR ([Banerjee & Lavie, 2005](#)), ROUGE ([Lin, 2004](#)), and CIDEr ([Vedantam et al., 2015](#)). In the context of Visual Question Answering (VQA), models are evaluated based on their ability to answer questions on an image's content. Here, the accuracy of model responses is compared against human-annotated answers, serving as the primary metric ([Antol et al., 2015](#)) to ensure alignment with human preferences.

However, when tackling sophisticated visual-language tasks, conventional evaluation metrics often fail to accurately capture the nuanced responses generated by these models, especially in complex or subjective scenarios that involve both visual elements and extended textual content ([Liu et al., 2023a](#)). Additionally, while manual annotation offers a more comprehensive and human-like evaluation, it comes with significant challenges. These include high costs ([Prendki, 2023](#)), potential biases ([Zheng et al., 2023b](#)), and the difficulty of ensuring consistent replication ([Chiang & Lee, 2023a](#)). These limitations highlight the need for a more holistic approach to evaluation, one that combines human-like calibration with more fine-grained assessment methods.

## B. Detailed Benchmark Construction

### B.1. Step 1: Image-Instruction Collection

To attain the outlined objectives, our approach begins with a detailed analysis of the capabilities of MLLMs. Specifically, we focus on the following abilities within MLLMs:

- **Recognition Ability:** This encompasses general visual recognition capabilities, such as object recognition, Optical

Character Recognition (OCR), and other high-level tasks in computer vision (Yu et al., 2023b).

- **Comprehension Ability:** This pertains to the model’s proficiency in spatial understanding and scenario comprehension.
- **Inferential Ability:** This involves the model’s capacity to process information and reasoning, a critical component in processing charts, graphs, and mathematics.
- **Multilingual Ability:** This assesses the model’s competence in understanding and processing multiple languages, especially focusing on their appearance in visual tasks such as text reading on images (Singh et al., 2019).

To ensure a robust and comprehensive assessment, we meticulously identify and incorporate 10 diverse datasets 8 into our evaluation framework. This strategic selection aims to enrich the diversity of our assessment tasks and enhance the breadth and depth of our evaluation capabilities, as well as prevent biases. These datasets are chosen based on their ability to effectively challenge the various aspects of MLLMs, via different downstream tasks, ensuring a thorough and nuanced understanding of their performance and potential.

To construct a robust and unbiased set of image-instruction pairs, we randomly select 300 images from each dataset, ensuring a diverse representation. Specifically, for the MathVista dataset, which includes the provision of hints, we extract 600 corresponding instructions, encompassing both scenarios: with and without hints. For the remaining datasets, we align 300 instructions with the sampled images. This process culminates in a comprehensive collection comprising 4,114 images corresponding with 4,414 instructions.

Table 8. Datasets and corresponding tasks in benchmark construction, each task is matched with several required abilities. (Rec.-Recognition, Comp.-Comprehension, Inf.-Inferential, Mul.-Multilingual)

Dataset	Image Type	#Images	#Questions	#Selected Pairs	Task	Ability Required
<b>Conceptual Captions</b> (Sharma et al., 2018)	Web image	3.3M	–	300	Captioning	Rec.&Comp.
<b>ChartQA</b> (Masry et al., 2022)	Chart	21K	33K	300	Chart reasoning	Rec.&Comp.
<b>InfographicVQA</b> (Mathew et al., 2021)	Infographics	5.4K	30K	300	Graph reasoning	Rec.&Comp.
<b>MathVista</b> (Lu et al., 2023)	Mathematics	6K	6K	300	Math reasoning	Rec.&Comp.&Inf.
<b>TextVQA</b> (Singh et al., 2019)	Text	28K	45K	300	Text reading	Rec.&Comp.
<b>WIT</b> (Srinivasan et al., 2021)	Multilingual text	11.5M	–	300	Transcription	Rec.&Mul.
<b>MS COCO</b> (Lin et al., 2014)	Real-life scene	328K	2.5M(labels)	300	Image Segmentation	Rec.&Comp.
<b>DiffusionDB</b> (Wang et al., 2022)	Diffusion	14M	1.8M(prompts)	300	Comprehensive	Rec.&Comp.&Inf.
<b>CC-3M Concept-balanced</b> (Liu et al., 2023d)	Comprehensive	595K	595K	300	Comprehensive	Rec.&Comp.&Inf.
<b>VisIT-Bench</b> (Bitton et al., 2023)	Comprehensive	1K	592	300	Instruction following	Rec.&Comp.&Inf.
<b>Mind2Web</b> (Deng et al., 2024)	Webpage	2K	2K	300	Website Understanding	Rec.&Comp.&Inf.
<b>AesBench</b> (Huang et al., 2024b)	Aesthetics Perception	3K	8K	300	Aesthetics Perception	Rec.&Comp.&Inf.
<b>ScienceQA</b> (Lu et al., 2022)	Science Knowledge	21K	21K	300	Reasoning	Comp.&Inf.
<b>MMvet</b> (Yu et al., 2023b)	Comprehensive	214	214	214	Instruction following	Rec.&Comp.&Inf.

**B.2. Step 2: MLLM Responses Collection**

We engage with 4 mainstream MLLMs (i.e., GPT-4V, Gemini, LLaVA, CogVLM) by providing them with our assembled image-instruction pairs for the first 3,300 image-instruction pairs, each VLM generated a response, resulting in a comprehensive collection of 13,200 answers, with each of the 3,300 instructions receiving a distinct response from each of the four MLLMs. For the last 4 datasets, we added during the rebuttal, we leverage GPT-4V, Gemini, Qwen-VL-Max, and

LLaVA-1.6-34b. For the sequential dataset Mementos (Wang et al., 2024b), we leverage GPT-4V, Qwen-VL-Max, ChatUnivi (Jin et al., 2023), VideoChat2 (Li et al., 2023b) to generate responses. Upon collecting a total of 17,656 responses from the MLLMs, we proceed to analyze the distribution of response lengths for each model. Figure 8 is a detailed illustration of length distribution in corresponding datasets.

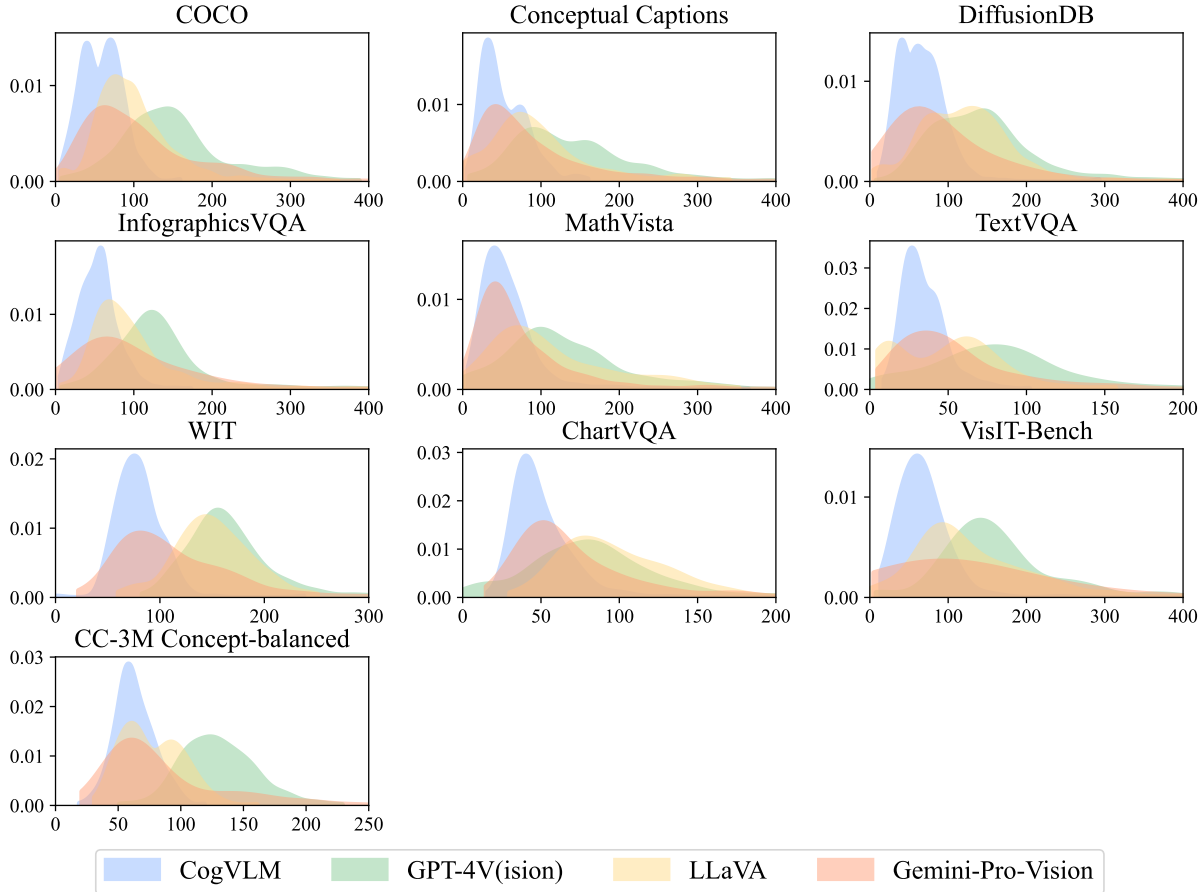


Figure 8. Response length distribution in 10 datasets. The horizontal axis represents length, and the vertical axis represents density.

## C. Detailed Experiment Settings

### C.1. Response VLM Settings

We use GPT-4V, Gemini, LLaVA-1.5-13b, CogVLM, Qwen-VL-Max, LLaVA-1.6-34b to answer the image-instruction pair. We discuss their hyperparameter settings and problems encountered during inference respectively:

- **GPT-4V** (OpenAI, 2023). We set the temperature and top-p as 0.9, max-token as 2048. However, we encounter some situations where it cannot answer accurately or refuses to answer due to ethical issues like *Unfortunately, due to my programming, I'm unable to ...*, which brings some difficulties to us in defining its judging capability.
- **Gemini** (GeminiTeam, 2023). We use the default settings, which set temperature as 0.4, top-p as 1, and max-token as 2048. It should be noted that Gemini will receive more ethical limitations than GPT-4V, and will refuse to answer on the diffusion data set. But for some more difficult questions, it can't answer the questions, but it will "forcibly answer" the user's questions. In this case, GPT-4V will sincerely admit its shortcomings and give a possible answer.
- **LLaVA-1.5-13b** (Liu et al., 2023d). We set temperature as 0, tok-p as 1, max-token as 2048, and beam search number as 3. The reason why we select such a low temperature is that LLaVA cannot correctly output its judge in a specific



format. We collect responses by inference on a dual-4090 local server.

- **CogVLM** (Wang et al., 2023c). For the hyper-parameter, we use the default setting and set max-token as 2048. We collect responses by inference on a dual-4090 local server.
- **Qwen-VL Family** (Bai et al., 2023a). We use the default settings for Qwen-VL family, with top-p as 0.8 and max-token as 2048.
- **LLaVA-1.6 Family** (Liu et al., 2023c). We set the temperature as 0.4 and top-p as 0.9, max-token as 2048.

## C.2. GPT-4V as Judge

We adopt GPT-4V without using JSON Mode mod based on our preliminary experiment in Appendix D.3, but required it to output in JSON format in our prompt. Following the hyper-parameter set in (Chiang & Lee, 2023b), we set the temperature to 0.9, top- $k$  to 0.9, and max-token to 2048 in both cases with and without pictures. When there is a format error in the output or ethical settings are triggered, we will sample again. If it is still wrong, we will skip this piece of data.

## C.3. Gemini-Vision-Pro as Judge

We call Gemini’s API on the Google Cloud Server and use the default settings for temperature and top- $k$ . It should be noted that even though Gemini is currently the only VLM that can perform Judge, it will occasionally speak Chinese or Japanese during the output process when using English text and English pictures, shown as Unicode in Figure 9. Although it is semantically continuous, that is, it can be used after translation, and it greatly affects the continuity of the judge. We believe that this is a capability issue, and it is speculated that it is affected by the multi-linguist in some training sets.

## C.4. Other MLLMs as Judge

We also use other mainstream MLLMs such as CogVLM (Wang et al., 2023c) and LLaVA (Liu et al., 2023d) for judging, which hyper-parameters follow the response generation settings. Some issues arise when we use the basic “Analyze-then-Judge” setting, where they cannot follow the judging instructions very well. Therefore we do not put their results into the results of the text. Recently, we have also noticed the advent of GLM-4V<sup>2</sup>, which has shown good performance on some benchmarks and can receive long-term text and follow the instructions for judging. However, due to the time constraints, we have not completed tests on GLM-4V in our work.

---

<sup>2</sup><https://open.bigmodel.cn/>



www.shutterstock.com · 626955464

**Instruction:** Here is the caption of this figure: human hands touch an alchemical circle. According to this figure and the caption, please give a detailed analysis of this figure:

**Answer:**

錬金術の円の上に人間の 手が触れています。錬金術の円は、円の周りに均等に配置された5つの同心円から構成されています。円の中心には、目、三日月、星、その他の記号が描かれています。人間の手は、円の両側にあり、手のひらが上を向いています。手のひらの指は、円の周りに描かれた記号と重なっています。この図は、錬金術師が錬金術の円を使って魔術を行っている様子を描いています。錬金術師は、円の周りに描かれた記号を使って、魔術的な力を呼び起こそうとしています。人間の手は、錬金術師の魔術的な力を表しています。

**Analysis:** The assistant’s response is in **Japanese**. It provides a detailed analysis of the figure, including the symbols on the alchemical circle and the possible meaning of the hands touching it. The response is well-written and easy to understand.”

**Judgment:** 5

Figure 9. Gemini speaks Japanese.

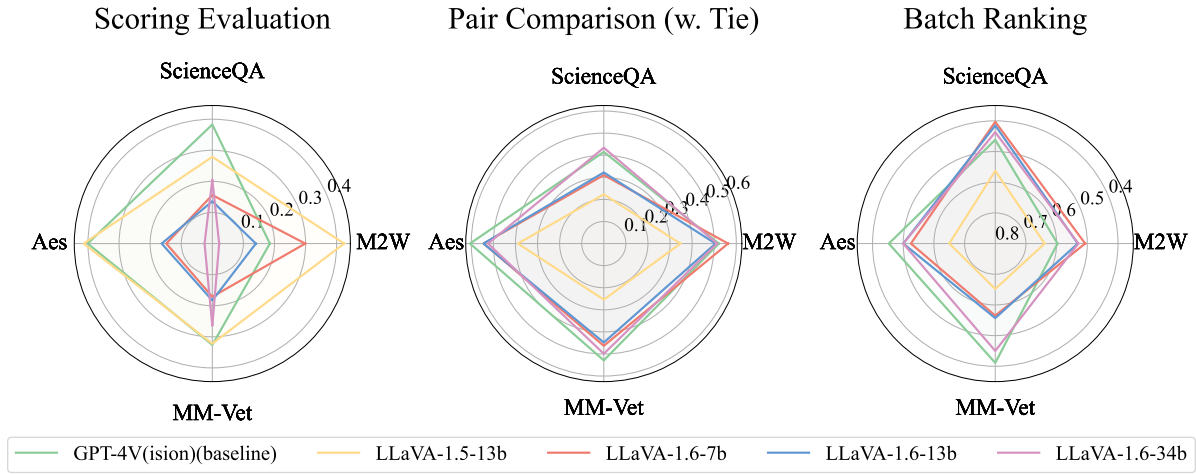


Figure 10. Comparative performance of different MLLMs across three judging settings in four newly added datasets, each is the average of three iterations.

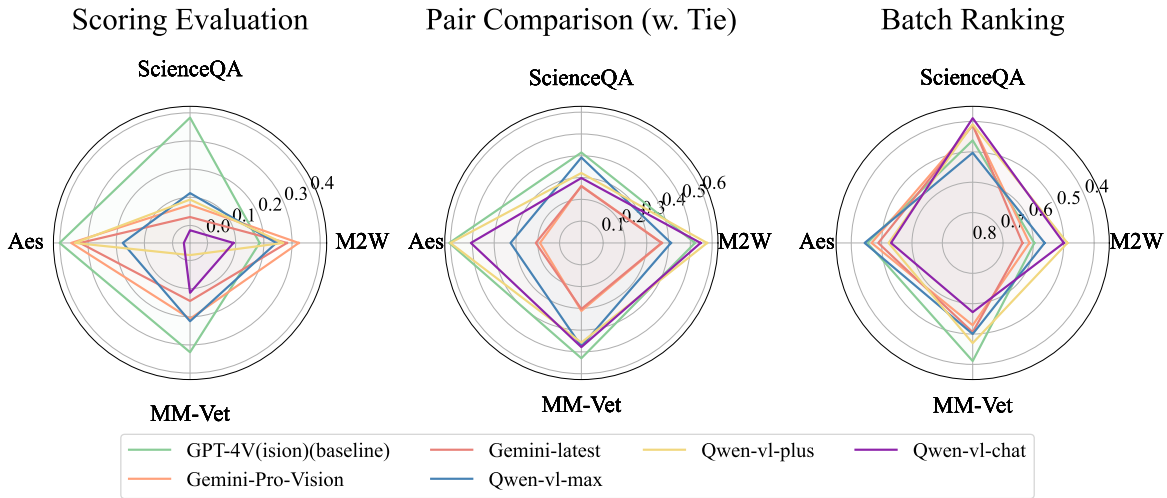


Figure 11. Comparative performance of different MLLMs across three judging settings in four newly added datasets, each is the average of three iterations.

## D. Additional Experimental Results

### D.1. Full Results on Judging Performance

We provide full results of judging performance of different MLLMs in Table 9. Comparative performance for four newly added datasets are shown in Figures 10 and 11.

In *Scoring Evaluation*, all models demonstrated comparable performance levels on the original dataset presented in our study, with LLaVA-1.6-34b and Qwen-vl-max slightly outperforming others in Math, Chart, and Text tasks, yet none surpassing GPT-4V. Our analysis of Qwen-vl-max and Qwen-vl-plus revealed a propensity to assign higher scores, with 80% of their ratings falling between 4-5 points, and a noticeable absence of 1-2 point scores. This inclination towards higher scores is more pronounced compared to other models. The LLaVA-1.6 series, although slightly better, also tends to award scores within the 3-5. In *Pair comparison*, qwen-vl-plus and max performed better on certain datasets, distinguishing themselves from competitors. Notably, qwen-vl-max exhibited less positional bias than LLaVA models, which showed a strong predisposition to favor one position, typically rating ‘A’ as better. In *Batch Ranking*, the updated Gemini-Pro-Vision model outperforms others overall. Both Qwen and LLaVA series demonstrated that larger model sizes correlate with better outcomes, affirming a strong scaling law effect. Despite these findings, there remains a noticeable gap between these models

and the top-performing GPT-4V, particularly concerning positional bias.

Table 9. The overall performance of different MLLMs in judging, compared with human annotations on different datasets. We sample all the data three times and took the average to mitigate the casualty. *w.* and *w.o.* tie represents tie and non-tie situations respectively. We omit Gemini’s results on the diffusion task for its challenges in processing AI-generated images. All presented data of Pearson similarity exhibit a *p*-value below 0.05, indicating a statistically significant level of confidence. Notice: Gemini-Pro\* means Gemini-1.0-Pro-latest.

Settings	MLLM	COCO	C.C.	Diff.	Graphics	Math	Text	WIT	Chart	VisIT	CC-3M	M2W	SciQA	Aes	MM-Vet	Ave.
Score (↑)	CogVLM	0.107	-0.048	0.049	-0.158	0.065	0.097	-0.131	-0.135	0.278	0.157	-	-	-	-	0.028
	GPT-4V	<b>0.454</b>	<b>0.507</b>	<b>0.458</b>	<b>0.645</b>	<b>0.606</b>	<b>0.624</b>	<b>0.579</b>	<b>0.645</b>	<b>0.620</b>	<b>0.431</b>	0.185	<b>0.383</b>	0.401	<b>0.326</b>	<b>0.490</b>
	LLaVA-1.5-13b	0.247	0.227	0.060	0.242	0.093	0.245	0.109	0.237	0.177	0.071	<b>0.424</b>	0.279	<b>0.414</b>	0.322	0.225
	LLaVA-1.6-7b	0.300	0.243	0.058	0.200	0.090	0.193	0.044	0.085	0.228	0.026	0.299	0.156	0.148	0.171	0.160
	LLaVA-1.6-13b	0.289	0.226	-0.110	0.078	0.056	0.086	0.062	0.120	0.163	0.200	0.140	0.136	0.163	0.183	0.128
	LLaVA-1.6-34b	0.285	0.251	-0.012	0.262	0.238	0.258	0.151	0.318	0.198	0.109	0.022	0.206	0.025	0.265	0.184
	Gemini-Pro	0.262	0.408	-	0.400	0.228	0.222	0.418	0.343	0.336	0.374	0.324	0.073	0.360	0.207	0.304
	Gemini-Pro*	0.211	0.230	0.114	0.146	0.060	0.095	0.041	0.160	0.174	0.177	0.282	0.030	0.329	0.144	0.157
	Qwen-vl-max	0.311	0.117	0.072	0.218	0.175	0.196	0.028	0.312	0.151	0.045	0.244	0.115	0.177	0.216	0.170
	Qwen-vl-plus	-0.050	0.195	0.019	0.126	0.106	0.161	0.151	0.089	0.128	0.106	0.268	0.092	0.347	-0.019	0.123
Qwen-vl-chat	-0.012	-0.012	0.033	-0.422	0.011	-0.028	0.021	0.036	-0.060	0.083	0.092	-0.017	-0.040	0.115	-0.014	
Pair w. Tie (↑)	CogVLM	0.548	0.409	0.562	0.613	0.412	0.250	0.273	0.262	0.324	0.433	-	-	-	-	0.409
	GPT-4V	<b>0.696</b>	<b>0.824</b>	<b>0.847</b>	<b>0.639</b>	0.564	<b>0.673</b>	<b>0.679</b>	<b>0.657</b>	0.640	<b>0.612</b>	0.521	0.415	<b>0.606</b>	<b>0.529</b>	<b>0.636</b>
	LLaVA-1.5-13b	0.273	0.478	0.286	0.273	<b>0.657</b>	0.510	0.369	0.383	0.456	0.484	0.347	0.223	0.389	0.254	0.384
	LLaVA-1.6-7b	0.493	0.571	0.550	0.383	0.314	0.507	0.500	0.352	0.401	0.402	0.563	0.310	0.544	0.463	0.454
	LLaVA-1.6-13b	0.493	0.586	0.590	0.333	0.339	0.507	0.587	0.296	0.454	0.459	0.506	0.322	0.545	0.448	0.462
	LLaVA-1.6-34b	0.493	0.600	0.570	0.300	0.374	0.551	0.543	0.254	0.398	0.392	0.513	<b>0.434</b>	0.524	0.499	0.460
	Gemini-Pro	0.616	0.787	-	0.650	0.436	0.664	0.605	0.500	<b>0.660</b>	0.560	0.370	0.262	0.190	0.312	0.509
	Gemini-Pro*	0.273	0.273	0.240	0.324	0.237	0.275	0.136	0.377	0.232	0.294	0.368	0.260	0.209	0.303	0.272
	Qwen-vl-max	0.403	0.464	0.372	0.494	0.438	0.500	0.533	0.479	0.421	0.421	0.411	0.392	0.325	0.474	0.438
	Qwen-vl-plus	0.479	0.507	0.650	0.450	0.328	0.522	0.500	0.380	0.453	0.383	<b>0.577</b>	0.321	0.601	0.457	0.472
Qwen-vl-chat	0.493	0.486	0.480	0.311	0.248	0.406	0.543	0.310	0.332	0.292	0.547	0.298	0.507	0.478	0.409	
Pair w.o. Tie (↑)	CogVLM	0.654	0.450	0.643	0.704	0.481	0.292	0.500	0.423	0.500	0.591	-	-	-	-	0.524
	GPT-4V	<b>0.804</b>	<b>0.870</b>	<b>0.922</b>	<b>0.807</b>	<b>0.801</b>	<b>0.805</b>	0.734	<b>0.849</b>	<b>0.761</b>	<b>0.703</b>	0.699	0.647	<b>0.755</b>	0.659	<b>0.773</b>
	LLaVA-1.5-13b	0.327	0.537	0.302	0.300	0.726	0.684	0.600	0.610	0.648	0.583	0.449	0.443	0.498	0.344	0.504
	LLaVA-1.6-7b	0.593	0.597	0.618	0.434	0.468	0.636	0.561	0.471	0.436	0.466	0.633	0.621	0.568	0.705	0.558
	LLaVA-1.6-13b	0.614	0.612	0.663	0.382	0.487	0.618	0.659	0.420	0.503	0.549	0.576	0.598	0.565	0.620	0.562
	LLaVA-1.6-34b	0.607	0.824	0.855	0.402	0.587	0.750	<b>0.758</b>	0.381	0.503	0.564	<b>0.712</b>	<b>0.679</b>	0.694	<b>0.762</b>	0.648
	Gemini-Pro	0.717	0.840	-	0.770	0.678	0.793	0.688	0.658	0.711	0.652	0.471	0.358	0.265	0.400	0.615
	Gemini-Pro*	0.311	0.340	0.308	0.419	0.336	0.366	0.200	0.439	0.290	0.358	0.469	0.336	0.266	0.398	0.345
	Qwen-vl-max	0.657	0.674	0.556	0.667	0.635	0.732	0.647	0.638	0.560	0.586	0.608	0.646	0.741	0.662	0.644
	Qwen-vl-plus	0.596	0.556	0.771	0.554	0.463	0.735	0.575	0.535	0.521	0.510	0.659	0.612	0.627	0.659	0.598
Qwen-vl-chat	0.603	0.523	0.625	0.333	0.386	0.574	0.625	0.431	0.370	0.396	0.618	0.594	0.539	0.755	0.527	
Batch (↓)	GPT-4V	<b>0.318</b>	0.353	<b>0.070</b>	<b>0.385</b>	<b>0.348</b>	<b>0.319</b>	<b>0.290</b>	<b>0.347</b>	<b>0.300</b>	0.402	0.597	0.462	0.453	<b>0.411</b>	<b>0.361</b>
	LLaVA-1.5-13b	0.577	0.492	0.562	0.535	0.598	0.650	0.616	0.644	0.620	0.563	0.639	0.563	0.650	0.652	0.597
	LLaVA-1.6-7b	0.575	0.538	0.618	0.462	0.601	0.598	0.564	0.679	0.586	0.503	0.507	0.403	0.525	0.565	0.552
	LLaVA-1.6-13b	0.614	0.612	0.663	0.382	0.487	0.618	0.659	0.420	0.503	0.549	0.531	0.415	0.500	0.557	0.536
	LLaVA-1.6-34b	0.449	0.411	0.500	0.561	0.575	0.544	0.483	0.552	0.542	0.479	0.529	0.437	0.500	0.450	0.501
	Gemini-Pro	0.287	<b>0.299</b>	-	0.473	0.462	0.430	0.344	0.520	0.426	<b>0.357</b>	0.613	0.412	0.467	0.529	0.432
	Gemini-Pro*	0.378	0.370	-	0.572	0.508	0.452	0.417	0.572	0.492	0.434	0.636	0.412	0.489	0.506	0.480
	Qwen-vl-max	0.477	0.407	0.500	0.480	0.507	0.515	0.493	0.539	0.468	0.407	0.563	0.503	<b>0.444</b>	0.500	0.486
	Qwen-vl-plus	0.640	0.616	0.500	0.666	0.644	0.634	0.592	0.747	0.671	0.540	<b>0.488</b>	0.409	0.523	0.470	0.581
	Qwen-vl-chat	0.733	0.701	0.500	0.669	0.638	0.554	0.638	0.723	0.687	0.668	0.500	<b>0.389</b>	0.531	0.572	0.607

## D.2. Judging Results on Sequential Images

We incorporated the sequential image dataset Mementos, comprising picture sequences, to expand our MLLM-as-a-Judge framework into the video domain in a pioneering effort. Each sequence, featuring over four images, draws from daily life, comics, and robotics. For data generation in Step 3, we utilized GPT-4V, Qwen-VL-Max, Qwen-VL-Plus, and Video-LLM ChatUnivi, obtaining 100 image-text pairs for batch evaluations, 381 for scoring, and 560 for pair comparisons. Beyond analyzing GPT-4V and Qwen-vl-max, we explored Video-LLM’s judging capabilities, specifically testing it with ChatUnivi.

As illustrated in Table 10 for *Batch Evaluation*, *Pair Comparison*, and *Score Evaluation* respectively, our findings indicate that GPT-4V significantly outperforms other models on sequential data. Despite the high-quality responses generated by the Video-LLM ChatUnivi we evaluated, it fundamentally lacks the judging capability and consistency.

Table 10. Judging performance on image sequence dataset Mementos.

MLLM	Score ( $\uparrow$ )	Pair ( $\uparrow$ )		Batch ( $\downarrow$ )
	Pearson	w. Tie	w.o. Tie	Edit Dis.
GPT-4V	<b>0.361</b>	<b>0.721</b>	<b>0.836</b>	<b>0.411</b>
ChatUnivi	-0.094	0.158	0.168	0.556
Qwen-vl-plus	0.115	0.426	0.482	0.5
Qwen-vl-max	0.046	0.446	0.531	0.63

### D.3. Preliminary Experiment

**Human Agreement on GPT-4V Output Mode.** The recently introduced ‘Json Mode’<sup>3</sup> in GPT-4V represents a significant advancement, particularly in structuring outputs in JSON format while restricting token usage. This mode has been observed to regularize responses, a feature particularly advantageous when dealing with structured data. However, this structuring tends to compartmentalize responses, potentially leading to a loss in the natural flow and contextual linkage typically inherent in human-like responses. This segmentation might inadvertently affect the readability and perceived coherence of the generated text.

To quantitatively assess the impact of Json Mode on output quality and its alignment with human preferences, we meticulously construct a test set. This set comprises 50 data instances, randomly selected from three distinct datasets used for evaluation purposes. The objective is to discern human evaluators’ predilection for the outputs generated in Json Mode by GPT-4V.

For a comprehensive analysis, we engage three annotators, each responsible for labeling the data. Their assessments aim to discern the balance between structured, JSON-formatted responses and the inherently fluid nature of human judgment and preference in textual content, as shown in Figure 12.

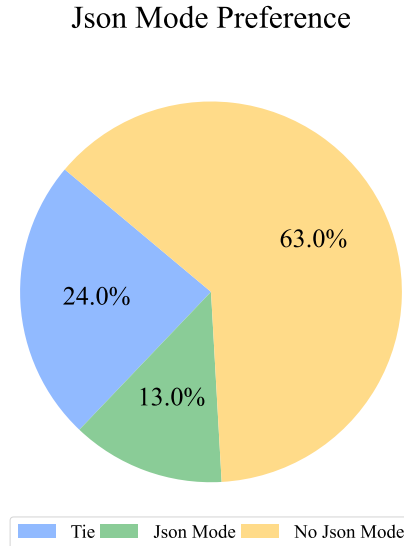


Figure 12. Json Mode Preference Analysis.

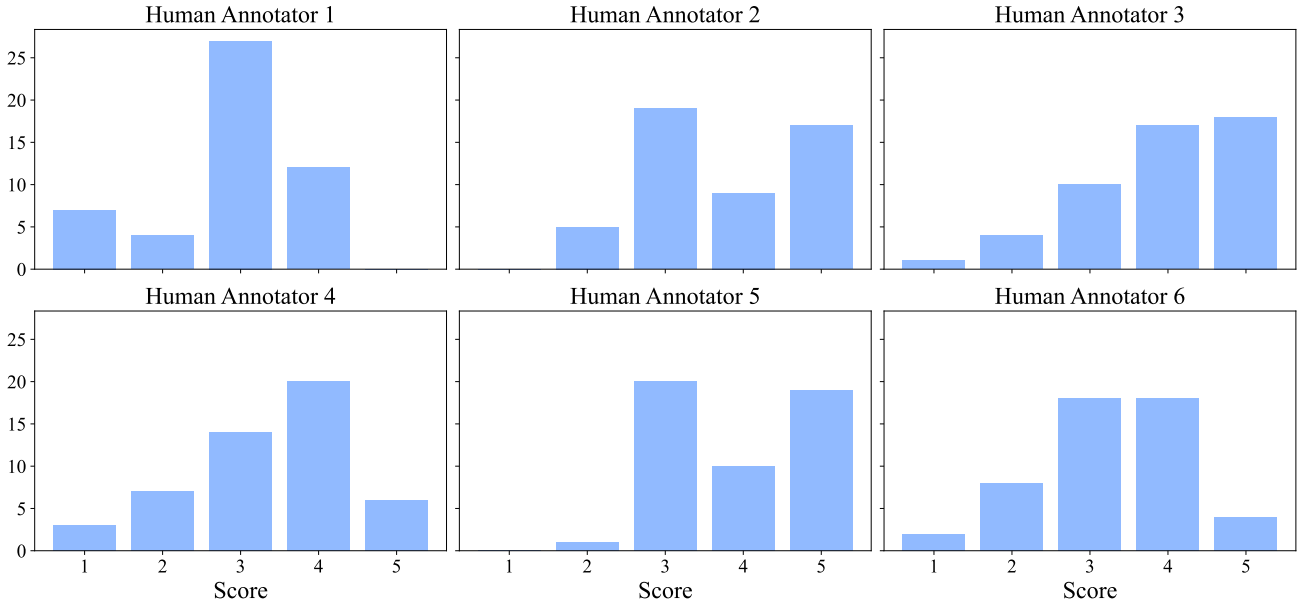
**Human Agreement Bias Checking** Acknowledging the inherent variability in human annotations, we embark on an empirical study involving ten annotators to ascertain the reliability of derived statistical patterns, notwithstanding the subjective nature of human judgment. This study aims to mitigate the individual biases that might skew the evaluation of GPT-4’s outputs. A dataset comprising 50 entries, processed using the GPT-4 pair comparison setting, serves as the foundation for this investigation.

The results, detailed in Figure 13, underscore a noteworthy observation: while the annotators exhibit minimal variance in determining the correctness of GPT-4’s judgments, a discernible divergence emerged in the scoring of analytical responses.

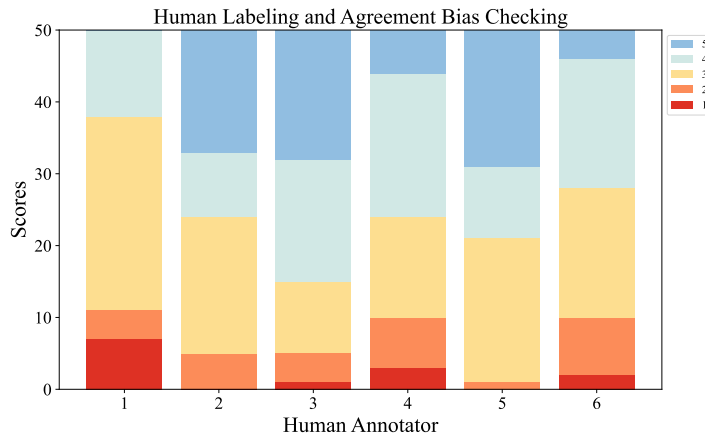
<sup>3</sup><https://openai.com/blog/new-models-and-developer-products-announced-at-devday>

This divergence presumably stems from individual perceptual differences and inherent biases. However, it’s crucial to note that these discrepancies in scoring did not significantly compromise the overall integrity of the annotations.

A remarkable consensus is observed in the labeling of hallucinations. The employment of a meticulously defined decision tree for identifying hallucinations ensures a high degree of uniformity across the annotations. This structured approach substantially minimizes errors, underscoring the effectiveness of well-defined criteria in achieving consistent and reliable annotations across different individuals.



(a) The distribution of Human Annotators’ ratings for the data.



(b) Human Labeling and Agreement Bias Checking.

Figure 13. Human Labeling and Agreement Bias.

#### D.4. Length Distribution on MLLM Judgments Analysis

In our analysis, we have included length distribution diagrams that showcase the differences in the responses provided by GPT-4V and Gemini during their judgment tasks as illustrated in Figure 14. These diagrams reveal that GPT-4V typically generates longer responses than Gemini in both *Scoring Evaluation* (Figure 15) and *Pair Comparison* (Figure 16), whereas in the batch task (Figure 17), the output lengths from both models are comparatively similar.

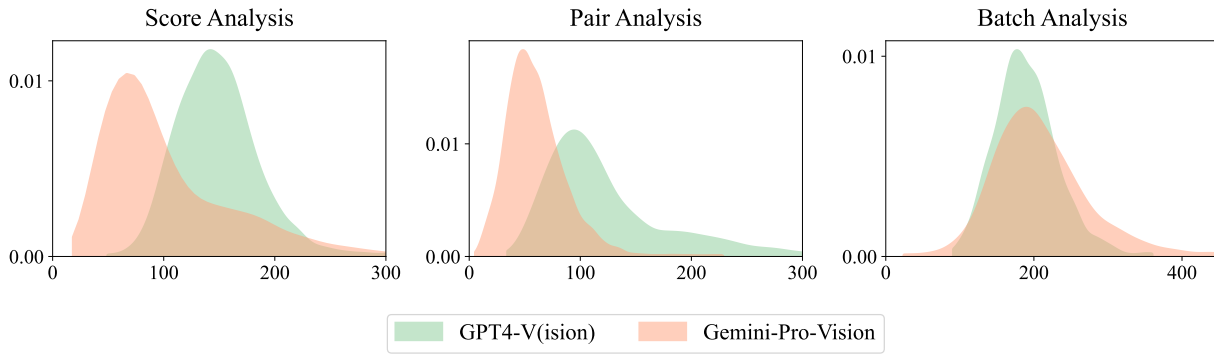


Figure 14. Length distribution in analysis collections.

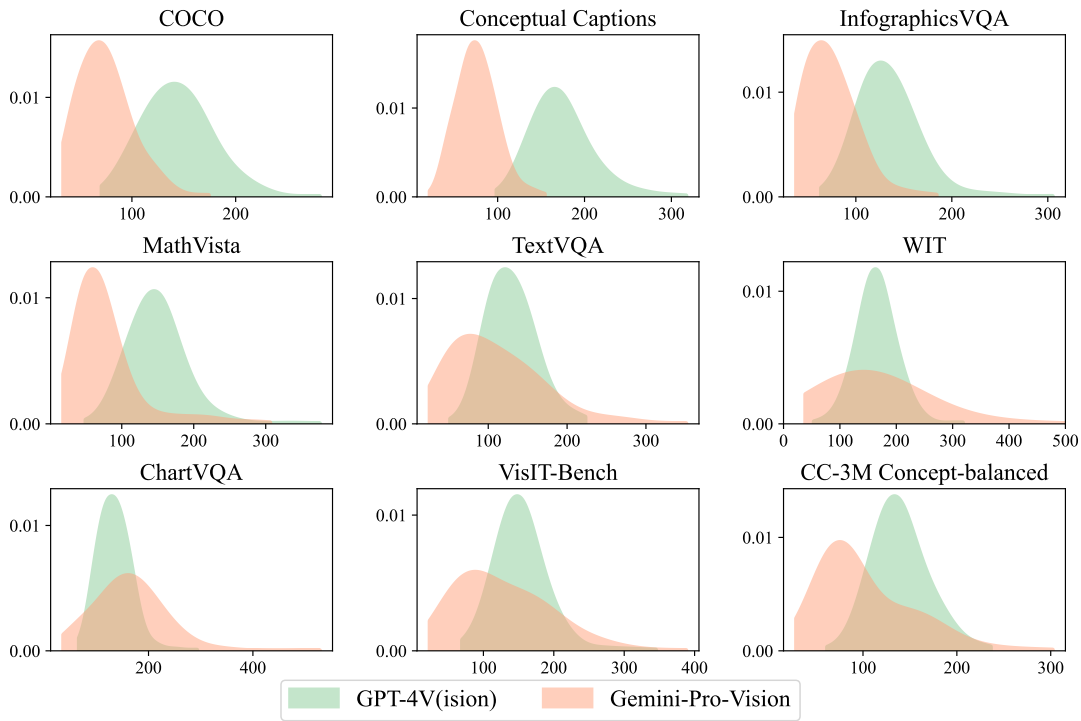


Figure 15. Response length distribution in Scoring Evaluation. The horizontal axis represents length, and the vertical axis represents density.

### D.5. Results on Human Scoring and Ego Bias

We employ the Mean Absolute Deviation (MAD) metric to assess the consistency of MLLM judging quality across multiple responses to a single image-instruction pair, as shown in 18.

The Egocentric Bias of different models are shown in Figures 19 and 20.

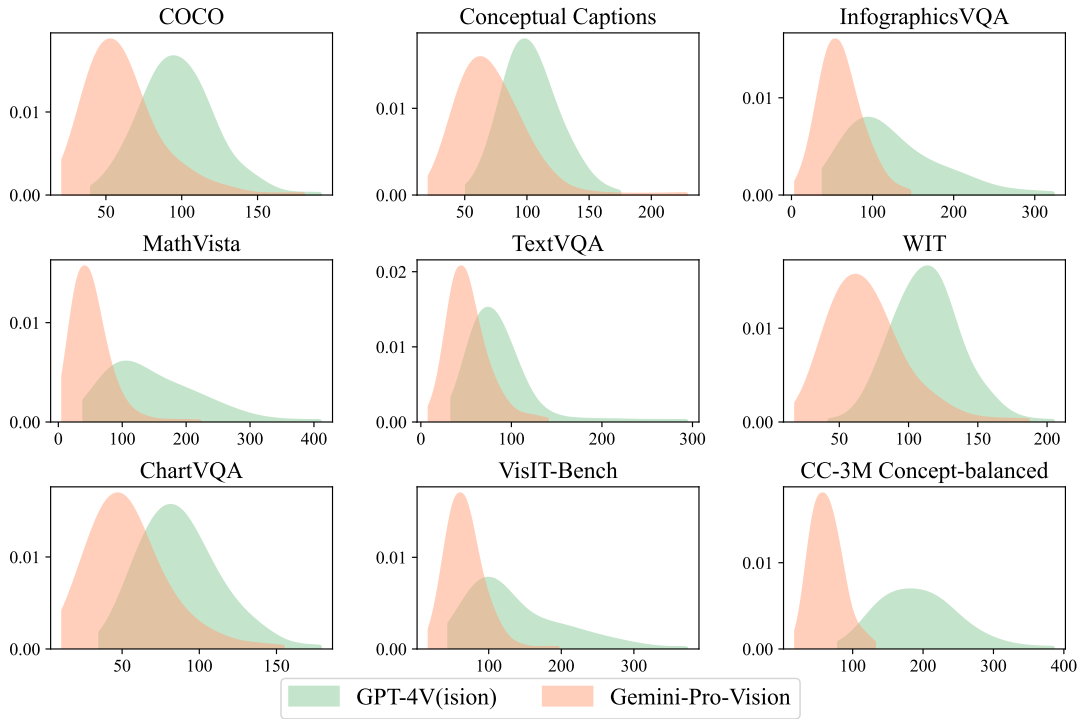


Figure 16. Response length distribution in Pair Comparison.

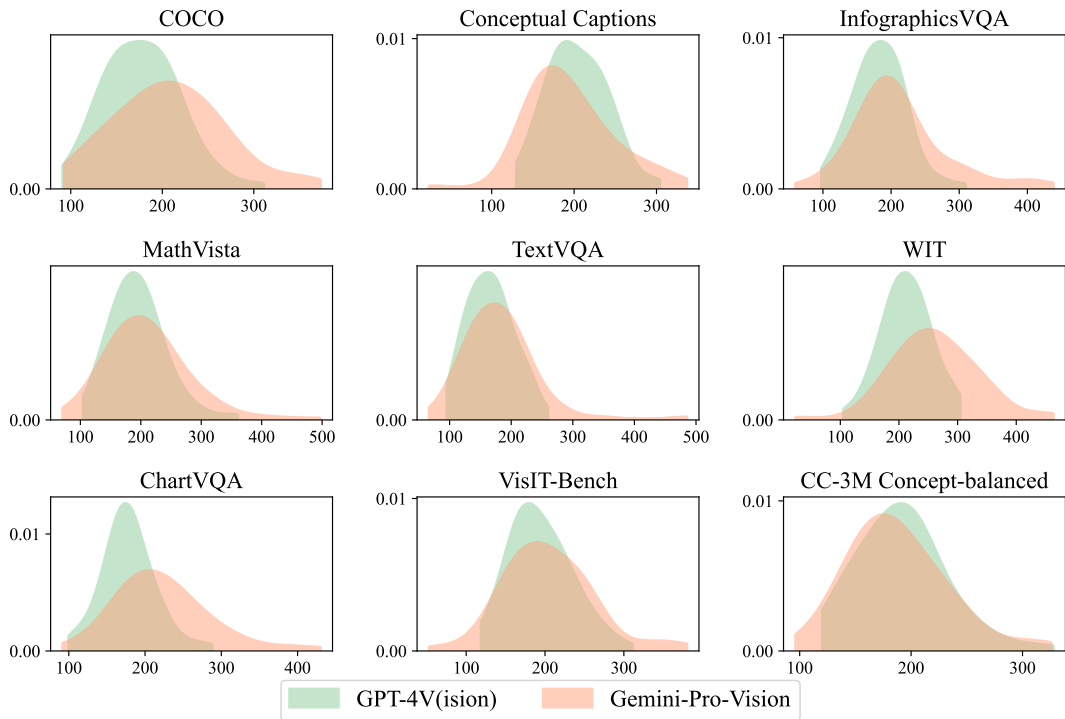


Figure 17. Response length distribution in Batch Ranking.



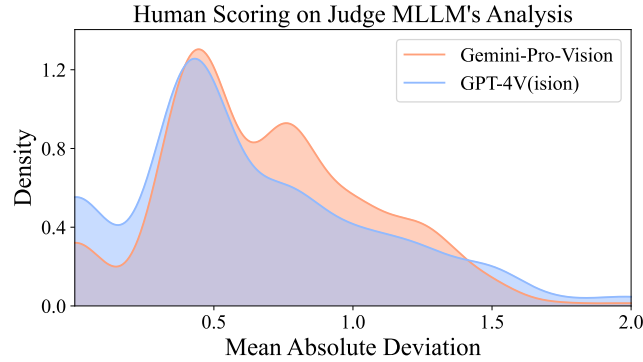


Figure 18. MAD of Human Scoring on MLLM Judgments Analysis.

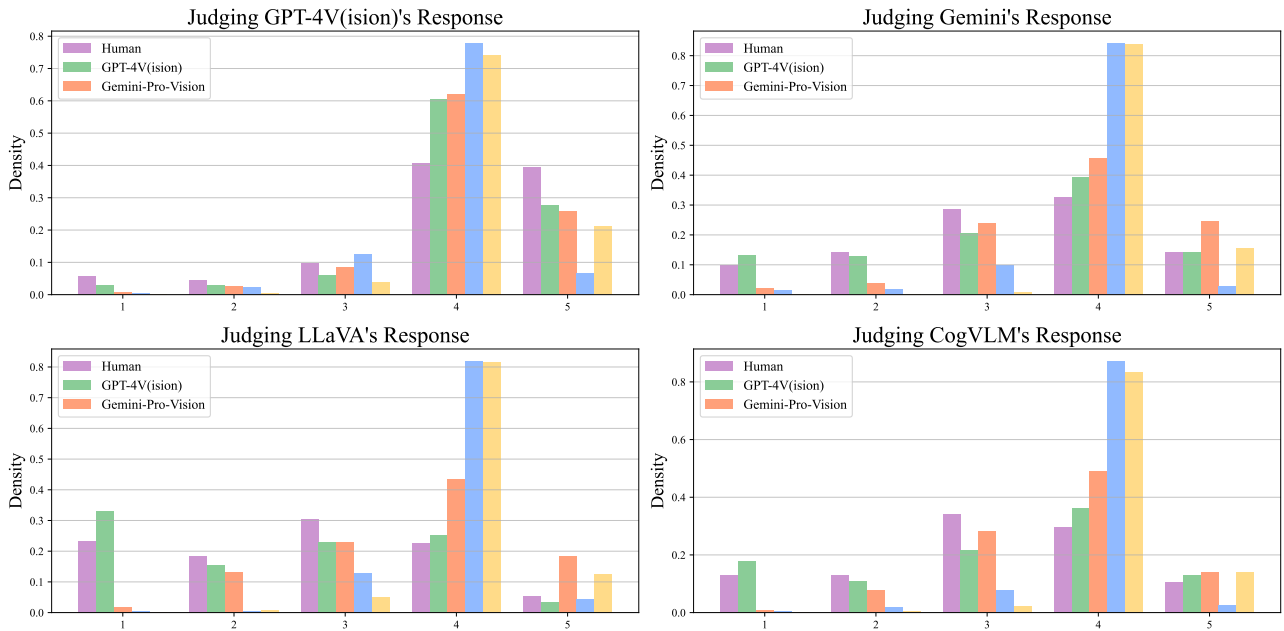


Figure 19. Scoring Density of Different MLLMs in Judging.

### E. Human Labeling and Agreement Collection

The annotation is conducted by 6 authors of this paper independently. As acknowledged, the diversity of annotators plays a crucial role in reducing bias and enhancing the reliability of the benchmark. These annotators have knowledge in this domain, with different genders, ages, and educational backgrounds. To ensure the annotators can proficiently mark the data, we provide them with detailed tutorials, teaching them how to evaluate model responses more objectively. Specifically, they are required to give judgments without considering answer lengths, and certain names or positions of the response. Besides, we implement cross-validation between different annotators and conduct continuous monitoring to ensure they are maintaining objectivity and fairness.

In the Human agreement experiment performed by humans on MLLM Judge, the prompt we give humans is presented in Figure 21 and Figure 22.

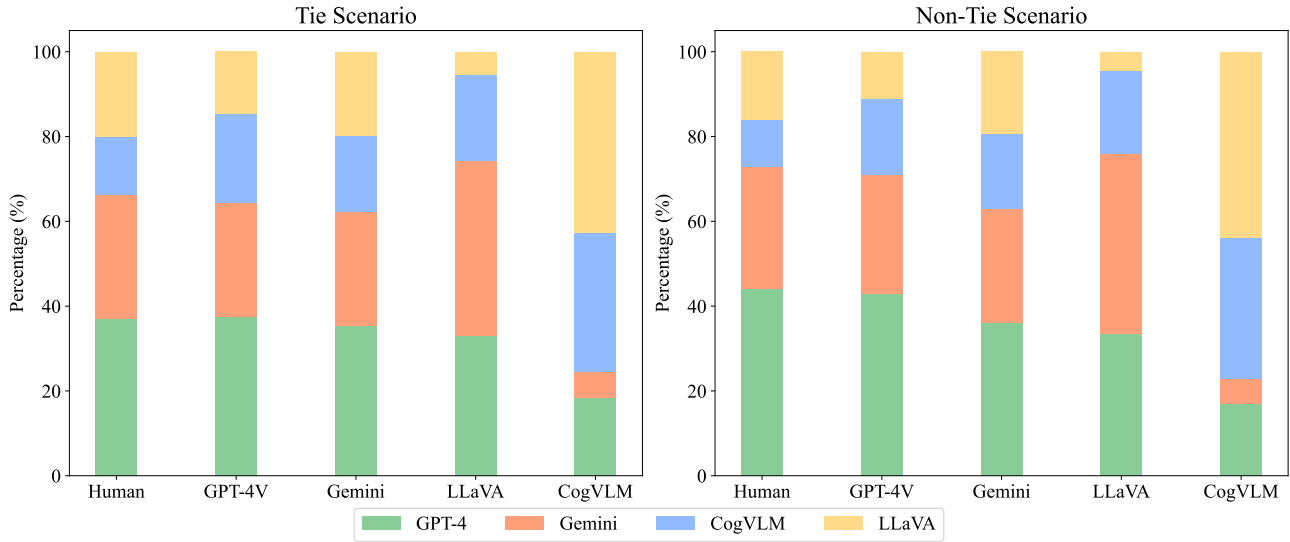


Figure 20. The proportion of different responses chosen by humans and different MLLMs in Tie Scenario and Non-Tie Scenario.

Prompts for Human Agreement Experiment

Your assessment should identify whether the assistant effectively adheres to the user’s instructions and addresses the user’s inquiry.

Do not allow the length of the responses to influence your evaluation.

Do not favor certain names or positions of the assistants. Be as objective as possible.

In your evaluation, weigh factors such as relevance, accuracy, comprehensiveness, creativity, and the granularity of the responses:

**Relevance:** The judge’s decisions directly correspond to the provided instructions or criteria. Every judgment made is pertinent to the case at hand, without deviation into unrelated areas.

**Accuracy:** The judge’s decisions are consistently in line with the established rules or guidelines. There is a clear understanding and correct application of these guidelines in every judgment.

**Comprehensiveness:** The judge considers all necessary aspects and evidence related to each case. Every relevant point in the guidelines is addressed in the judge’s evaluation.

**Creativity:** The judge demonstrates the ability to approach complex or ambiguous situations with innovative thinking. This includes providing insightful, constructive feedback or solutions not explicitly covered in the guidelines.

**Granularity of Responses:** The judge offers detailed and specific reasoning for each decision. This entails a thorough breakdown of how each aspect of the guidelines applies to the case or situation at hand.

Prompt for human

Judgement of the result: Whether you agree with the judgement of the result. Explain if you agree with the analysis of the result and provide reasons.  
Analysis rating of the result: Rate the quality of the result's analysis on a scale of 1 to 5, where 1 is very dissatisfied and 5 is very satisfied.

- 1 point: The analysis is completely inaccurate, with significant missing or incorrect information.
- 2 points: The analysis is partially accurate, but there are several errors or important information is missing.
- 3 points: The analysis is mostly accurate, but there are a few errors or missing information.
- 4 points: The analysis is quite accurate, with only very minor errors or omissions.
- 5 points: The analysis is very accurate, complete with information, and virtually error-free.

Hallucination: Judge whether the analysis of the result exhibits hallucination. Hallucination refers to the presence of information in the analysis that does not correspond to reality or is completely fabricated.

Comparison: Judge which result is better. Compare the analyses of different results, point out which one is more accurate, more complete with information, and provide reasons for your choice.



Instruction

This is a figure including mathematical information. Please analyze this figure in detail and answer the following question with hint based on this figure: Hint: Please answer the question and provide the correct option letter, e.g., A, B, C, D, at the end. Question: Is this nest larger than a fist? Choices: (A) Yes (B) No

Please read through result1.judge and result1.analysis

result1.judge: B  
result1.analysis: Assistant A provides more detailed reasoning and arrives at the correct answer. Assistant B's answer is short and less convincing.

Do you agree with the judgment of result1?

yes  no

Rate the analysis of result1

1  2  3  4  5

Is there any hallucination in the analysis of result1?

yes  no

Figure 21. Human agreement

Prompt for human

Your assessment should identify whether the assistant effectively adheres to the user's instructions and addresses the user's inquiry. Do not allow the length of the responses to influence your evaluation. Do not favor certain names or positions of the assistants. Be as objective as possible. In your evaluation, weigh factors such as relevance, accuracy, comprehensiveness, creativity, and the granularity of the responses:

- Relevance: The judge's decisions directly correspond to the provided instructions or criteria. Every judgment made is pertinent to the case at hand, without deviation into unrelated areas.
- Accuracy: The judge's decisions are consistently in line with the established rules or guidelines. There is a clear understanding and correct application of these guidelines in every judgment.
- Comprehensiveness: The judge considers all necessary aspects and evidence related to each case. Every relevant point in the guidelines is addressed in the judge's evaluation.
- Creativity: The judge demonstrates the ability to approach complex or ambiguous situations with innovative thinking. This includes providing insightful, constructive feedback or solutions not explicitly covered in the guidelines.
- Granularity of Responses: The judge offers detailed and specific reasoning for each decision. This entails a thorough breakdown of how each aspect of the guidelines applies to the case or situation at hand.



Instruction

This is a task of text reading on natural image. Please analyze this figure in detail and answer the following question with reason based on this figure: what country was this coin made in?

Please read through ResponseA

The coin was made in the United States of America. This can be determined by reading the inscription 'UNITED STATES OF AMERICA' around the edge of the coin, which explicitly states the country of origin. Additionally, the coin features the head of 'Liberty', a figure frequently represented on U.S. currency, and the year '1907', which is consistent with historic U.S. coinage.

Please read through ResponseB

The coin in the image is a United States of America dime. This can be inferred from the inscription "United States of America" and the year "1907" on the coin's surface. The dime also features an image of a Native American, which is a common design element on American coins. The presence of these inscriptions and the image of the Native American on the coin indicate that it was made in the United States.

Which answer do you think is better?

- ResponseA is better     ResponseB is better     Equally good

Figure 22. Human labeling

## F. Prompt Templates

We first query Judge VLM to get their responses by the following prompts.

### Query prompts of MLLMs in judging.

You are a helpful assistant proficient in analyzing vision reasoning problems.

[The Start of User Instruction]

{item['instruction']}

[The End of User Instruction]

Please provide a detailed explanation for your response.

Following [Chiang & Lee \(2023b\)](#) and [Li et al. \(2024\)](#), we have designed prompts and presented the prompt template of VLM’s operation including score, pair comparison, and batch ranking judgments in a prompt template as *system prompt*, *instruction*, *criteria*, *noticement*, and *desired output form*:

### Template prompts of scoring evaluation

#### (System Prompt)

You are a helpful assistant proficient in analyzing vision reasoning problems.

#### (Instruction)

Please examine the provided image attentively and serve as an unbiased judge in assessing the quality of the response from an AI assistants regarding the instruction. You will receive a single response from the assistant to user’s instruction.

#### (Noticement)

Your assessment should identify whether the assistant effectively adheres to the user’s instructions and addresses the user’s inquiry.

In your evaluation, weigh factors such as relevance, accuracy, comprehensiveness, creativity, and the granularity of the responses.

Do not allow the length of the responses to influence your evaluation.

Do not favor certain names or positions of the assistants. Be as objective as possible.

#### (Criteria)

Use scores to show the quality of the response. Here is the detailed scoring rubric for evaluating the quality of responses from AI assistants:

Poor (1): The response significantly deviates from the user’s instruction and fails to address the query effectively. It shows a lack of relevance, accuracy, and comprehensiveness. Creativity and granularity are absent or poorly executed.

Fair (2): The response addresses the user’s instruction partially, with evident shortcomings in relevance, accuracy, or comprehensiveness. It lacks depth in creativity and granularity, indicating a superficial understanding of the user’s inquiry.

Average (3): The response adequately addresses the user’s instruction, showing a fair level of relevance, accuracy, and comprehensiveness. It reflects a basic level of creativity and granularity but may lack sophistication or depth in fully capturing the user’s inquiry.

Good (4): The response is well-aligned with the user’s instruction, demonstrating a high degree of relevance, accuracy, and comprehensiveness. It shows creativity and a nuanced understanding of the topic, with a detailed granularity that enhances the response quality.

Excellent (5): The response perfectly adheres to the user’s instruction, excelling in relevance, accuracy, comprehensiveness, creativity, and granularity. It provides an insightful, detailed, and thorough answer, indicating a deep and nuanced understanding of the user’s inquiry.

#### (Desired Output Format)

Use "[[1]]", "[[2]]", "[[3]]", "[[4]]", "[[5]]" to indicate your evaluate score in the key ‘Judgement’.

[The Start of User Instruction]

{item['instruction']}

[The End of User Instruction]

[The Start of Assistant’s Answer]

{item['answer']}

[The End of Assistant’s Answer]

## Template prompts of pair comparison

**(System Prompt)**

You are a helpful assistant proficient in analyzing vision reasoning problems.

**(Instruction)**

Please examine the provided image attentively and serve as an unbiased judge in assessing the quality of responses from two AI assistants regarding the user's question shown beneath the image.

**(Noticement)**

Your assessment should identify the assistant that more effectively adheres to the user's instruction and aptly addresses the user's inquiry.

In your evaluation, weigh factors such as relevance, accuracy, comprehensiveness, creativity, and the granularity of the responses.

Avoid any position biases and ensure that the order in which the responses were presented does not influence your decision.

Do not allow the length of the responses to influence your evaluation.

Do not favor certain names of the assistants. Be as objective as possible.

Present your verdict in a JSON format, with the key 'analysis' for a short reason of your judgement and the key 'judgment' to indicate your decision: use "[[A]]" if assistant A prevails, "[[B]]" if assistant B does, and "[[C]]" for a tie.

**(Desired Output Format)**

[The Start of User Instruction].

```
{item['instruction']}
```

[The End of User Instruction]

[The Start of Assistant A's Answer]

```
{item['answer1']['answer']}
```

[The End of Assistant A's Answer]

[The Start of Assistant B's Answer]

```
{item['answer2']['answer']}
```

[The End of Assistant B's Answer]

## Template prompts of batch ranking

**(System Prompt)**

You are a helpful assistant proficient in analyzing vision reasoning problems.

**(Instruction)**

Please serve as an unbiased judge in assessing the quality of the responses from AI assistants regarding the user's instruction and a figure.

You will be presented with several responses from different assistants to the same user instruction. Your task is to assess and compare these responses based on how effectively they adhere to the user's original instruction and how aptly they address the user's inquiry. After your assessment and comparison, you should RANK the responses from best to worst as the following template. If Assistant A is the best response, Assistant D is the worst response, you should output like "[[A]], [[B]], [[C]], [[D]]"

**(Noticement)**

Your assessment should identify whether the assistant effectively adheres to the user's instructions and addresses the user's inquiry.

In your evaluation, weigh factors such as relevance, accuracy, comprehensiveness, creativity, and the granularity of the responses.

Do not allow the length of the responses to influence your evaluation.

Do not favor certain names or positions of the assistants. Be as objective as possible.

Indicate your final rank in the key 'Judgement'.

**(Desired Output Format)**

[The Start of User Instruction].

```
{item['instruction']}
```

[The End of User Instruction]

## G. Case Study

We list several case studies. Assistant A, B, C, D represent LLaVA, CogVLM, GPT-4V and Gemini, respectively. Judge VLM 1 and 2 represent Gemini and GPT-4V, respectively. For pair comparison, the judge result “C” means a tie.

An example of score comparison



**Instruction:** Please analyze this figure in detail and answer the following question based on this figure: Who was the train built for?

**answer1.name:** Assistant A

**answer1.answer:** The train was built for children to play with in a park. The image shows a small red and black toy train parked on a dirt surface in a park, surrounded by trees. The train appears to be designed for children to ride on and enjoy, making it an ideal attraction for families visiting the park.

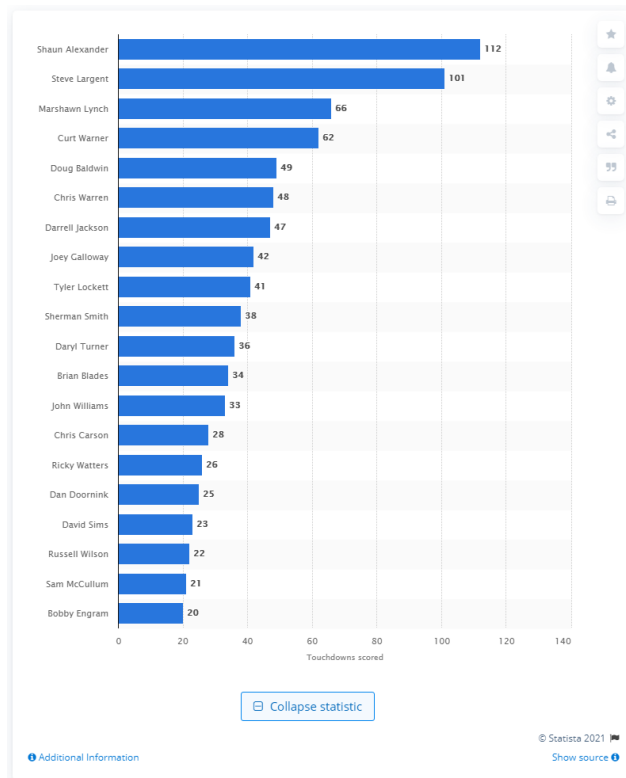
**result1-1.name:** Judge VLM 1

**result1-1.analysis:** The assistant’s response is relevant to the user’s instruction and addresses the user’s inquiry. The response is accurate and comprehensive, providing a clear and concise answer to the question. The response is also creative and uses detailed granularity to enhance the quality of the response. **Judgement: 5**

**result1-2.name:** Judge VLM 2

**result1-2.analysis:** The response provided by the AI assistant correctly identifies the purpose of the toy train in the image, stating that it was built for children to play with in a park. The description of the train as small, red and black, and parked on a dirt surface surrounded by trees is accurate based on the visual evidence in the image. Furthermore, the assistant correctly deduces that the train is designed for children to ride on and asserts that it is an ideal attraction for families, which is a logical conclusion given the context of the setting. The response is relevant to the user’s instruction and demonstrates accuracy and a reasonable level of comprehensiveness. There is a minor display of creativity in the assistant’s description of the train being an ideal attraction for families. **Judgement: 4**

## An example of pair comparison



**Instruction:** This is a task of reasoning on chart data. Please analyze this figure in detail and answer the following question with reason based on this figure. Who is the career touchdown leader of the Seattle Seahawks?

**answer1.name:** Assistant A

**answer1.answer:** The career touchdown leader of the Seattle Seahawks is Shaun Alexander with 112 touchdowns as illustrated by the bar chart, where his bar is the longest and the numerical value of 112 is the highest among all the listed players.

**answer2.name:** Assistant B

**answer2.answer:** The answer is Shaun Alexander. Shaun Alexander has 112 career touchdowns, the most of any player on the Seattle Seahawks. Steve Largent is second with 101 career touchdowns.

**result1-1.name:** Judge VLM 1

**result1-1.judge:** C

**result1-1.analysis:** Both assistants correctly answer the question and provide the same answer with sufficient supporting details.

**result1-2.name:** Judge VLM 2

**result1-2.judge:** B

**result1-2.analysis:** Both Assistant A and Assistant B correctly identified Shaun Alexander as the career touchdown leader for the Seattle Seahawks with 112 touchdowns, as shown in the bar chart. Assistant B provided additional context by mentioning that Steve Largent is second with 101 career touchdowns, offering a comparative perspective. While Assistant A's answer is accurate, Assistant B's answer adds a layer of detail that enriches the response without detracting from its relevance or accuracy.



An example of batch ranking



**Instruction:** This is a task of text reading on natural image. Please analyze this figure in detail and answer the following question with reason based on this figure. what year is at the top of the magazine?

**answer1.name:** Assistant A

**answer1.answer:** The year at the top of the magazine is 2014.

**answer2.name:** Assistant B

**answer2.answer:** The year at the top of the magazine is 2014. It is prominently displayed in large, bold letters, making it easily readable and the focal point of the image.

**answer3.name:** Assistant C

**answer3.answer:** The year at the top of the magazine is 2014. This can be determined by reading the text that says '9 Maggio – 10 Giugno 2014,' which indicates that the magazine or event was from May 9 to June 10, 2014. This text is found in the upper section of the image, in a red-colored font, just above the image of the two monsters.

**answer4.name:** Assistant D

**answer4.answer:** The year at the top of the magazine is 2014. The top of the magazine says "9 MAGGIO - 1 GIUGNO 2014". "9 MAGGIO - 1 GIUGNO" means "May 9 - June 1". "2014" is the year.

**result1-1.name:** Judge VLM 1

**result1-1.analysis:** Assistant A: The assistant provides a concise and straightforward response, stating that the year at the top of the magazine is 2014. However, the response lacks any further explanation or supporting details. Assistant B: Similar to Assistant A, this assistant also briefly states that the year is 2014, but it adds a bit more detail by mentioning that the year is displayed prominently in large, bold letters, making it easily readable and the focal point of the image. Assistant C: Assistant C offers a more comprehensive response by not only stating the year as 2014 but also providing a clear explanation. The assistant points out the specific text on the magazine that indicates the date range of May 9 to June 10, 2014, and correctly concludes that the year at the top is 2014. Assistant D: Assistant D's response is comparable to that of Assistant C, as it includes both the identification of the year as 2014 and an explanation. However, Assistant D's explanation is slightly different, as it mentions the Italian translation of the date range, which is 9 MAGGIO - 1 GIUGNO 2014. The response also includes additional information about the event, mentioning that it took place in Milan and was

organized by WOW Spazio Fumetto.

**Judgement** : Based on the analysis, I rank the responses as follows: [[C]], [[D]], [[B]], [[A]] Assistant C provides the most comprehensive and detailed response, clearly addressing the user's instruction and inquiry by identifying the year and providing supporting evidence from the image. Assistant D also offers a thorough explanation, along with additional information about the event. Assistant B's response is concise but lacks the level of detail provided by Assistants C and D. Assistant A's response is the briefest and lacks any explanation or supporting information.

**result1-2.name**: Judge VLM 2

**result1-2.analysis**: Assistant A: The response correctly identifies the year at the top of the magazine as 2014. However, it provides no additional information or context to support the answer., Assistant B: This response also correctly identifies the year as 2014 and goes further to describe the appearance of the text, which supports the user's instruction for detail and provides reasoning for the answer., Assistant C: The response is accurate in identifying the year as 2014 and provides a detailed description of where the date is located and what the text around it indicates. It adds context by mentioning the text color and the image below the date, giving a more comprehensive understanding., Assistant D: Assistant D accurately identifies the year as 2014 and gives a translation of the other text around the year, providing additional context. However, it slightly misquotes the text, stating '1 GIUGNO' instead of '10 Giugno,' which could be seen as a minor error in accurately reading the text. **Judgement**: Rank: [[C], [B], [A], [D]]