AlignMMBench: Evaluating Chinese Multimodal Alignment in Large Vision-Language Models

Anonymous ACL submission

Abstract

Evaluating the alignment capabilities of large Vision-Language Models (VLMs) is essential for determining their effectiveness as helpful assistants. However, existing benchmarks primarily focus on basic abilities using nonverbal methods, such as yes-no and multiple-choice questions. In this paper, we address this gap by introducing AlignMMBench, which provides more nuanced evaluations of alignment capabilities and is the first benchmark specifically designed for Chinese visual contexts. This benchmark is meticulously curated from real-world scenarios and internet sources, encompassing thirteen specific tasks across three categories, and includes both single-turn and multi-turn dialogue scenarios. Incorporating a prompt rewrite strategy, AlignMMBench encompasses 1,054 images and 4,978 question-answer pairs. To facilitate the evaluation pipeline, we propose CritiqueVLM, a rule-calibrated evaluator that exceeds GPT-4's evaluation ability. Additionally, we propose the "alignment score", a quantitative metric designed to assess the robustness and stability of models across diverse prompts. Finally, we report the performance of representative VLMs on AlignMMBench, offering insights into the capabilities and limitations of different VLM architectures. All evaluation codes and data are available on github[1].

1 Introduction

Equipped with the Large Language Models (LLMs), Vision-Language Models (VLMs) demonstrate impressive performance in various visual tasks, such as image description and visual question answering (Achiam et al., 2023; Wang et al., 2023; Bai et al., 2023a). Following alignment training, like Supervised Fine-Tuning (SFT) and Reinforcement Learning from Human Feedback (RLHF), VLMs can comprehend and generate human language within visual contexts.

Recently, significant efforts have been made into developing VLMs for not only English but also Chinese contexts, such as QwenVL (Bai et al., 2023a), CogVLM (Wang et al., 2023), MiniCPM (Hu et al., 2023), Intern-VL (Chen et al., 2023b), Intern-XComposer2V (Dong et al., 2024), Deepseek-VL (Lu et al., 2024a), and Yi-VL (AI et al., 2024). Specifically, in publicly available comprehensive rankings[2], these VLMs have demonstrated performance comparable to that of GPT-4o (Achiam et al., 2023). To facilitate objective comparisons and evaluations among these models, researchers mainly utilize the following benchmarks: 1) general evaluation sets such as MME (Fu et al., 2024) and MM-Bench (Liu et al., 2023a), and 2) domain-specific evaluation sets such as MathVista (Lu et al., 2024b) and MMMU (Yue et al., 2023). However, these benchmarks primarily assess model capabilities through non-verbal methods and lack detailed evaluations of alignment performance, particularly in the Chinese context.

However, constructing a high-quality Chinese multimodal alignment benchmark is fully challenging. First, unlike the English context, which benefits from traditional benchmarks such as VQAv2 (Goyal et al., 2017), TextVQA (Singh et al., 2019), and ChartQA (Masry et al., 2022), clean and publicly available Chinese multimodal corpora are exceedingly scarce. It is necessary to collect and construct these resources from scratch. Second, the inherent ambiguity of the Chinese context is more pronounced (Huang et al., 1997), necessitating multiple annotators to repeatedly verify a single high-quality Chinese multimodal corpus. Besides, The characteristics of images and the underlying world knowledge can vary significantly between different languages (Duncum, 2004). As a result, relying solely on English datasets limits the com-

[1]https://anonymous.4open.science/r/AlignMMBench-B689

[2]https://rank.opencompass.org.cn/leaderboard-multimodal

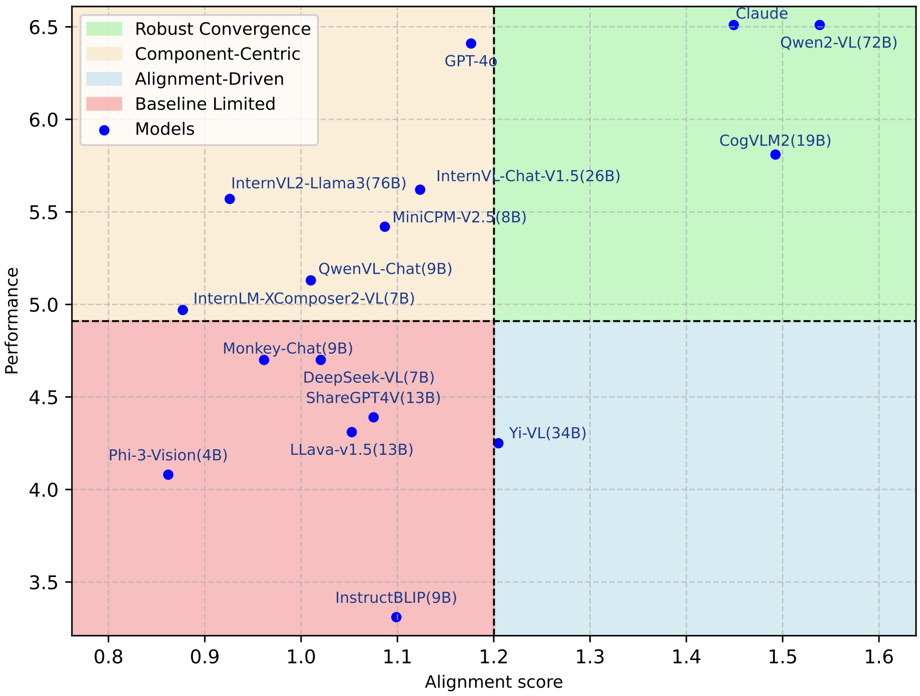


Figure 1: Performance vs. Alignment score of various models. Performances are ranged from 0 to 10, while Alignment scores are ranged from 0.2 to ∞.

prehensive evaluation of Chinese vision-language models (VLMs). Therefore, developing a multimodal benchmark within Chinese visual context is essential and promising.

To address these gap, we introduce AlignMMBench, a comprehensive evaluation benchmark specifically designed to assess the Chinese alignment capabilities. Given the scarcity of Chinese multimodal corpora, AlignMMBench is meticulously curated from real-world scenarios and internet resources. It encompasses thirteen specific tasks across three high-level categories, including both single-turn and multi-turn dialogue scenarios, to make detailed evaluation of alignment performance. Considering the diversity of query forms in real-world user scenarios, we introduce a LLM-based prompt-rewritting strategy, which transforms a single query into multiple stylistically distinct yet semantically equivalent questions. Leveraging this strategy, we introduce a new metric, the "alignment score", to investigate the reasons for performance differences among various models. By combining performance metrics with the alignment score, we categorize models into 4 classes, as deipcted in Figure 1. As shown in Table 1 and Figure 2, AlignMMBench comprises 1,054 images and 4,978 question-answer pairs.

As the lack of standard answers, the evaluation of open-ended questions is a significant and challenging research topic. One simple approach is to utilize the strongest LLM, such as GPT-4 (Achiam et al., 2023), to score the model responses. However, the API-based GPT-4 operates as a black box, resulting in limited control for developers. The costs associated with API usage impose limitations on the size of the datasets that can be used as well as the number of models that can be tested. To address these limitations, we fine-tuned an evaluator based on the ChatGLM3-6B (Du et al., 2022), named CritiqueVLM. Equipped with the detailed prompt rules and chain-of-thought reasoning, CritiqueVLM achieves a 34.8% reduction in mean absolute error when compared to human scoring, outperforming GPT-4. Additionally, CritiqueVLM features higher consistency across other metrics.

Based on AlignMMBench and CritiqueVLM,we evaluate numerous popular VLMs that support Chinese. Our findings are as follows: (1) All VLMs excel in perception and understanding, achieving an average score of 5.07, but perform poorly in reasoning and analysis, with an average score of 4.38; (2) Within dialogue contexts, VLMs exhibit suboptimal performance in the incoherence task compared to the coherence task, indicating that VLMs strug-

Benchmark	Dataset					Evaluation	
	Size	Language	Multi Category	Dialogue Context	Open Ended	Judge Method	Metric
Ai2D (Hiippala et al., 2021)	4,903	English	✗	✗	✓	multi-choice	Accuracy
LLaVABench (Liu et al., 2024)	150	English	✓	✗	✓	GPT	Score
MathVista (Lu et al., 2024b)	6,141	English[3]	✗	✗	✓	GPT	Accuracy
MME (Fu et al., 2024)	2,800	English	✓	✗	✗	yes-no	Accuracy
MMBench (Liu et al., 2023a)	3,217	Eng. & Chi.	✓	✗	✗	multi-choice	Accuracy
MMMU (Yue et al., 2023)	11,500	English	✗	✗	✗	multi-choice	Accuracy
MMStar (Chen et al., 2024b)	1,500	English	✓	✗	✗	multi-choice	Accuracy
MM-Vet (Yu et al., 2023b)	205	English	✓	✗	✓	GPT	Score
OCRBench (Liu et al., 2023b)	1,000	English	✗	✗	✓	string match	Accuracy
TouchStone (Bai et al., 2023b)	908	English	✓	✗	✓	GPT	Score
VisIT-Bench (Bitton et al., 2023)	592	English	✓	✗	✓	GPT	Elo / Win
AlignMMBench (ours)	**4,978**	**Chinese**	✓	✓	✓	**CritiqueVLM**	**Score**

Table 1: Comparisons between AlignMMBench and other benchmarks.

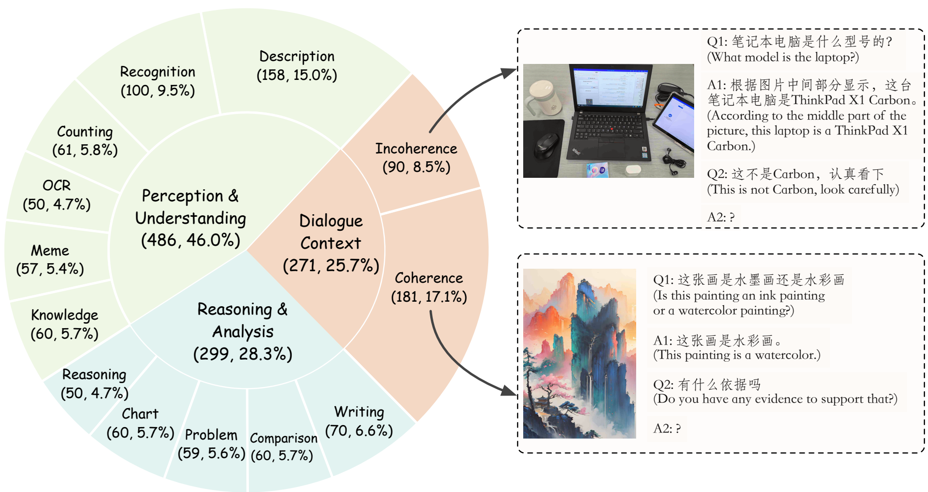


Figure 2: Categories and examples of AlignMMBench. The chart on the left displays the categories of AlignMMBench, encompassing three main categories and thirteen specific tasks. The numbers listed under each category represent the number of images in that category and the corresponding percentage of the total. The right side of the pie chart presents two examples, illustrating instances from the incoherence and coherence tasks.

gle to detect previous errors; (3) Certain English-based VLMs, such as Phi-3-Vision (Abdin et al., 2024), exhibit suboptimal performance on AlignMMBench, suggesting the composition of the training corpus is critical in alignment evaluation.

In summary, the main contributions of this paper are as follows:

- We propose AlignMMBench, a multimodal alignment benchmark that encompasses both single-turn and multi-turn dialogue scenarios. It includes three categories and thirteen capability tasks, with a total of 4,978 question-answer pairs. As far as we know, this is the first public alignment benchmark specifically designed for the Chinese visual context.
- To improve the controllability of evaluation scores, we introduce the CritiqueVLM, a ChatGLM3-6B based evaluator that has been rule-calibrated and carefully fine-tuned. With human judgements, its evaluation consistency surpasses that of GPT-4.
- We benchmark representative VLMs on AlignMMBench. Beyond their performance, we provide in-depth insights into the current state of Chinese VLM and highlight the areas that require further enhancement.

2 Related work

Multimodal benchmark. Early multimodal benchmarks predominantly focus on specific cross-modal tasks such as Image Caption (Lin et al., 2014; Plummer et al., 2015), Visual Grounding (Kazemzadeh et al., 2014; Yu et al., 2016), Visual Question Answering (Schwenk et al., 2022; Marino et al., 2019; Mathew et al., 2021) and Optical Character Recognition (OCR) (Singh et al., 2021). Recent benchmarks can be categorized into two types based on their questions, including domain-specific and general scenarios. 1) Domain-specific benchmarks concentrate on particular sub-fields. For example, discipline reasoning (Yue et al., 2023; Lu et al., 2024b), OCR (Liu et al., 2023b), chart understanding (Masry et al., 2022; Hiippala et al., 2021), and the hallucination test (Liu et al., 2024). 2) General benchmarks provide multi-dimensional and diverse assessments, encompassing MME (Fu et al., 2024), MMBench (Liu et al., 2023a), MMStar (Chen et al., 2024b), MMVet (Yu et al., 2023b), LLaVABench (Chen et al., 2024b), VisIT-Bench (Bitton et al., 2023), and Touch-Stone (Bai et al., 2023b). From another perspective, based on differences in result evaluation methods, the aforementioned benchmarks can be divided into deterministic and open-ended types. Table 1 provides an overview of the categories and salient features of prominent benchmarks.

Different from the above benchmarks, AlignMMBench is a universal and open-ended alignment benchmark. Additionally, we present dialogue context tasks based on real-world user corpus.

LLM-as-a-Judge methods. In the realm of automated model response evaluation, employing a robust LLM as an evaluation assistant is widely adopted across various VLM benchmarks. Math-Vista (Lu et al., 2024b) utilizes an LLM to extract answers from responses for more precise comparisons with reference answers, while other works (Liu et al., 2024; Yu et al., 2023b; Bai et al., 2023b; Bitton et al., 2023) adopt the LLM-as-a-Judge (Zheng et al., 2024) approach. This method prompts the LLM to act as a judge, scoring responses or selecting the better response between two candidates. Furthermore, recent researches (Ge et al., 2024; Zhang et al., 2023; Kim et al., 2024, 2023) explore the feasibility of employing the LVLM-as-a-Judge approach to achieve more accurate evaluation results.

However, evaluating numerous models within a large benchmark can be costly and inefficient, with results potentially becoming unstable due to API-based model updates. The current capabilities of LVLMs in instruction-following and logical reasoning are markedly inferior to those of the most advanced LLMs. In this work, we design a set of task-level evaluation prompts to improve the accuracy of LLM scoring. Additionally, we fine-tune ChatGLM3-6B (Du et al., 2022) using a high-quality, human-annotated grading corpus to enhance the evaluation capabilities of our model.

3 AlignMMBench

3.1 Dataset composition

AlignMMBench comprises 1,056 images and 4,978 associated test cases. Each test case includes an image, a question, and a meticulously crafted reference answer. To evaluate the capabilities of VLMs across various dimensions, we categorized these test cases into three primary categories and thirteen distinct tasks, ranging from simple object recognition and description to complex reasoning and computation, as shown in Figure 2. AlignMMBench is consist of three major categories: **Perception and understanding**, which involves answering questions by synthesizing information from images and world knowledge; **Reasoning and analysis** focuses on assessing the model's capabilities in information processing and analysis, which often require numerical calculations or logical reasoning to provide accurate responses; **Dialogue context** evaluates capability in real-world user scenarios. For a detailed definition on these categories and their sub-tasks, please refer to Appendix A.1.

3.2 Dataset construction

As illustrated Figure 3 (left), the construction of AlignMMBench involves three steps: image collection, query generation, and answer annotation.

Image collection. First, we manually define 13 task types in AlignMMBench, with researchers constructing detailed descriptions and instructions for each type. Subsequently, web crawlers were employed to retrieve images from Chinese websites, such as Baidu, based on these predefined descriptions. To ensure data quality, researchers manually filtered out low-quality or irrelevant images from the dataset. Then, we collected real-world user queries from a Chinese application that offers generative AI assistant services, primarily catering to professionals and students. Throughout our process, we were diligent in avoiding the extraction of images from websites with copyright restrictions, and we conducted manual inspections to ensure adherence to copyright compliance.

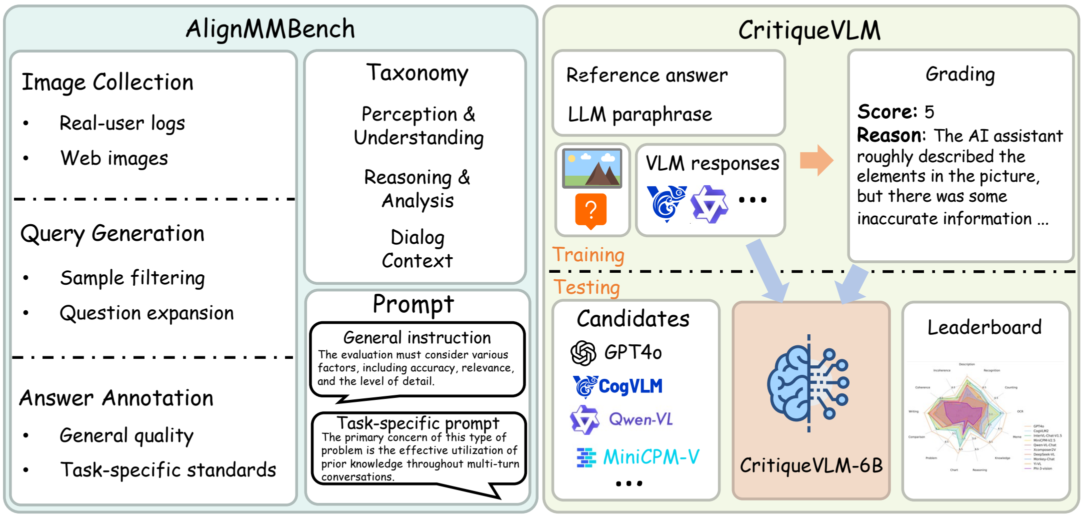


Figure 3: Overall framework of our work.

Furthermore, we manually eliminate low-quality images based on the following protocols: (1) Removing images containing unidentifiable objects or text; (2) Eliminating images that contains personal privacy or offensive content; and (3) Excluding images with similar content that have already been included. Besides, we calculated the MD5 value of each image to avoid inclusion in our known SFT datasets or prior benchmarks.

Query generation. For images obtained via the web crawler, we craft a seed question that correlates with the image and aligns with its designated category. Given the complexity of constructing a dialogue history based on an image, we use real-world user requests exclusively for building the "Dialogue Context" category. Considering the variability of problems with the same user intention, we employ ChatGLM (Du et al., 2022) to rephrase the seed questions without altering their original purpose. This method was applied specifically to single-turn questions.

Answer annotation. Annotators are instructed to produce accurate and comprehensive answers, incorporating as many pertinent details as possible. This approach enhances the ability to conduct precise evaluations independently of visual content.

Since we employ a prompt rewrite strategy to expand our question set, we implement a two-phase check process. In the first phase, each seed questions-answer pair (1,054 pairs in total) is manually reviewed. Subsequently, we utilize an LLM to expand the questions (details in Appendix A.4) and manually verify that each generated question is stylistically different yet semantically equivalent to seed questions. Following these checks, we obtained a total of 4,978 question-answer pairs.

Then, we manually eliminate low-quality responses based on the following protocols: (1) **Factual errors:** The reference answer includes a description that does not correspond to the image content; (2) **Unfollowing:** The reference answer does not follow with the instruction in the prompt; (3) **Incomplete answers:** The reference answer lacks sufficient information to evaluate the response without the image; (4) **Common sense mistakes:** The reference answer contains common sense errors.

Finally, we employed a two-phase annotation process to obtain our responses. In the first turn, annotators craft responses based on the corresponding images. After that, researchers review these preliminary annotations, filter out low-quality entries, and request re-annotations where necessary.

These steps ensure a robust foundation for the subsequent phases of our study, providing a reliable and high-quality multimodal benchmark.

4 CritiqueVLM

In this section, we propose an LLM-based evaluator, referred to as CritiqueVLM, to automatically assess responses based on the provided questions and detailed reference answers. CritiqueVLM is fine-tuned from ChatGLM3-6B (Du et al., 2022; Zeng et al., 2022) and generates a score between 1 and 10, along with a chain-of-thought explanation, as illustrated in Figure 4.

4.1 Training data construction

Prompt design. Following (Zheng et al., 2024), we devise a prompt system consisting of general

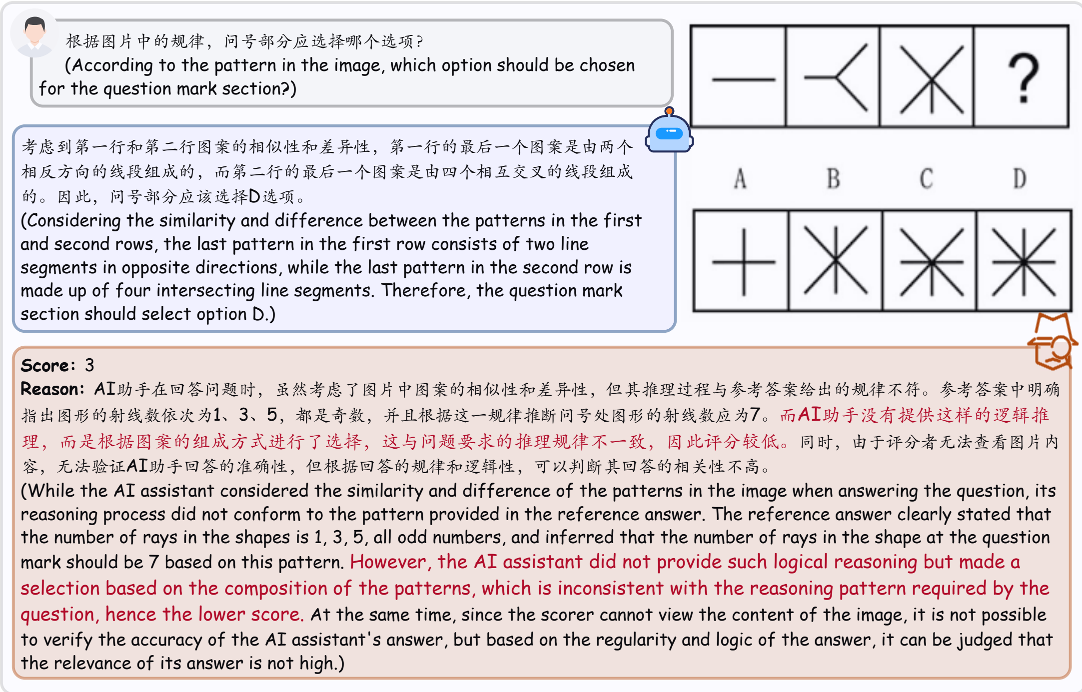


Figure 4: A critical example from CritiqueVLM.

and task-specific prompts. The general prompt outlines the scoring range, criteria, and output format. Task-specific prompts are varied by different categories, containing several hints to guide the critic model in accurately checking answers. Besides, we use in-context examples to mention that the critic model cannot access an image, encouraging it to infer the image's content. The detailed prompts are presented in Appendix A.8.

Scoring data collection. We collect the training corpus based on CogVLM (Wang et al., 2023) and QwenVL-Chat (Bai et al., 2023a). Given the challenging nature of the test cases in our AlignMMBench even for human annotators, we utilize responses from a "reference model" that rephrased reference answers using GLM-4 (Du et al., 2022; Zeng et al., 2022). This approach mitigates bias from predominantly low scores. Human annotators were then tasked with scoring each response from 1 to 10 and providing reasonable explanations, guided by the same prompts used by CritiqueVLM, but with access to images, questions, and reference answers. To prevent bias from the models' capabilities, we anonymize the model names and randomly shuffled the data.

Dataset partition. We allocate 90% of the data from each category to the training dataset and reserve the remaining 10% as an internal testing dataset. To assess the model's agreement with humans when analyzing responses from a novel VLM, we create an external testing dataset using responses generated by GPT-4 (Achiam et al., 2023) and annotated scores provided by human experts.

4.2 Supervised fine-tuning

We select ChatGLM3-6B (Du et al., 2022; Zeng et al., 2022) as the base model and fine-tune it on the training dataset. The DeepSpeed (Rasley et al., 2020) library was employed as our training framework, utilizing the FusedEmaAdam optimizer with a weight decay of 0.05. The learning rate is initially set to 10^{-5} with a 1% warmup ratio. The maximum source sequence length is set to 4096, and the target sequence length to 1024. Training is conducted with a global batch size of 128 on 32 A800 GPUs. After 1,000 iterations, the training loss decreased from 3.8 to 0.3.

5 Experiments

5.1 Baselines

We select two models, **ChatGPT** (gpt-3.5-turbo) and **GPT-4** (gpt-4-1106-preview)(Achiam et al., 2023), as baselines for CritiqueVLM, as they are widely utilized in other benchmark studies. Additionally, given that AlignMMBench is a Chinese-style benchmark, **ChatGLM3-6B**(Du et al., 2022; Zeng et al., 2022) is also included.

5.2 Evaluation on evaluator performance

In our experiment, we employ six statistical metrics to assess the agreement between human annotated

Testcase	Internal						External					
Metric	e	r	ρ	τ	f	s	e	r	ρ	τ	f	s
ChatGLM3-6B	2.424	0.230	0.224	0.194	0.350	0.285	3.237	0.103	0.091	0.077	0.296	0.197
ChatGPT	1.720	0.572	0.596	0.505	0.427	0.347	2.473	0.404	0.429	0.356	0.370	0.247
GPT-4	1.256	0.839	0.836	0.726	0.677	0.565	1.486	0.770	0.765	0.648	0.550	0.424
CritiqueVLM*	**0.818**	**0.846**	**0.838**	**0.740**	**0.747**	**0.646**	**1.146**	**0.778**	**0.782**	**0.671**	**0.670**	**0.511**

Table 2: Results of the agreement between human annotated scores and models.

scores and model generated scores. Common four statistical metrics include the mean absolute error (e), Pearson (r), Spearman (ρ), and Kendall correlation coefficient (τ). To mitigate the bias from annotators preferences, we map scores to predefined ranges and calculate accuracy as an evaluation metric. Based on scoring criteria, integers from 1 to 10 are divided into two sets of ranges:

- **Fuzzy division.** This division includes 4 ranges: Unfollow ($[1, 2]$), Bad ($[3, 5]$), Good ($[6, 8]$), and Excellent ($[9, 10]$). We denote the proportion of the score from annotation and model in the same ranges as f.
- **Strict division.** This division includes 7 ranges, each corresponding to specific ranges defined in the scoring criteria: $[1, 1], [2, 2], [3, 3], [4, 5], [6, 6], [7, 8], [9, 10]$. We denote the proportion of the score from annotation and model in the same ranges as s.

Detailed results are presented in Table 2, demonstrating that CritiqueVLM achieves superior performance with only 6 billion parameters. Additionally, Figure 4 provides two examples of scoring by CritiqueVLM, validating its ability to generate reasonable critical scores and detailed explanations.

5.3 Leaderboard

We benchmark a range of popular VLMs, including GPT4o (OpenAI., 2024), Qwen2-VL (Wang et al., 2024), CogVLM2 (Wang et al., 2023), InternVL2, InternVL-Chat-V1.5 (Chen et al., 2023b, 2024d,c; Gao et al., 2024), MiniCPM-V2.5 (Yu et al., 2023a; Hu et al., 2023; Xu et al., 2024; Yu et al., 2024), Qwen-VL-Chat (Bai et al., 2023a), XComposer2V (Dong et al., 2024), DeepSeek-VL (Lu et al., 2024a), Monkey-Chat (Li et al., 2023), Yi-VL (AI et al., 2024), Phi-3-Vision (Abdin et al., 2024), ShareGPT4V (Chen et al., 2023a), LLava-v1.5 (Liu et al., 2024), and InstructBLIP (Dai et al., 2023). Results are shown in Table 2.

5.3.1 Analysis on overall result.

The average scores of VLMs in Table 3 range from 3.3 to 6.5, indicating that most VLMs can understand question requirements and generate responses relevant to the images, according to our scoring criteria. Scores below 5.0 reflect numerous errors in the VLMs' responses. Additionally, GPT4o (OpenAI., 2024) demonstrates the best performance across most tasks, and CogVLM2 (Wang et al., 2023) secures the second-best performance.

Moreover, the disparity in rankings between the "Ref." column and AlignMMBench underscores the limitations of existing benchmarks, which do not fully account for characteristics in Chinese context. Consequently, AlignMMBench serves as a valuable complement to existing benchmarks.

5.3.2 Analysis on alignment ability.

We propose a novel metric, "alignment score", to evaluate the alignment capabilities of LVLMs. Consider a dataset with N seed questions, where each seed question, denoted as S_i, generates a set of M_i semantically equivalent questions $\{Q_{i,1}, Q_{i,2} \dots, Q_{i,M_i}\}$. For each question $Q_{i,j}$, let $R_{i,j}$ represent the score achieved by the model on this question. The proposed metric is defined as the inverse of the average standard deviation across all seed questions and is expressed as follows:

$$\text{Align.} = \frac{N}{\sum_{i=1}^{N} \sigma_i}$$

$$\sigma_i = \sqrt{\frac{1}{M_i} \sum_{j=1}^{M_i} (R_{i,j} - \overline{R}_i)^2}$$

$$\overline{R}_i = \frac{1}{M_i} \sum_{j=1}^{M_i} R_{i,j}$$

This metric reflects the average variability of results within clusters of semantically equivalent questions, with higher values indicating greater consistency. We agree that a well-aligned model can demonstrate consistent performance when presented with stylistically distinct yet semantically equivalent questions. Accordingly, this metric is selected as an evaluation criterion to assess the model's alignment capabilities.

By combining the average score and the alignment score, these models can be categorized into four groups, as illustrated in Figure 1:

Models	Size	Ref.	Avg	Perception & Understanding.						Reasoning & Analysis					Context		Align.
				Des.	Rec.	Cou.	OCR.	Mem.	Kno.	Rea.	Cha.	Pro.	Com.	Wri.	Coh.	Inc.	
Qwen2-VL	72B	1	**6.51**	7.39	6.64	6.64	**7.60**	7.09	**6.32**	4.00	7.16	**5.89**	6.57	7.72	6.37	5.26	**1.54**
Claude	-	4	**6.51**	7.68	**6.89**	**6.79**	7.02	7.10	6.28	4.06	7.11	5.20	5.92	**7.98**	**7.02**	**5.52**	1.45
GPT-4o	-	2	6.41	**7.75**	6.41	5.20	7.17	**7.28**	6.16	**4.44**	**7.23**	5.81	**7.19**	7.85	6.41	4.43	1.18
Cogvlm2	19B	8	5.81	7.20	6.12	5.75	7.21	6.07	5.69	3.43	5.92	4.37	5.65	7.34	6.33	4.43	1.49
InternVL-Chat	26B	5	5.62	7.12	6.00	5.51	6.63	4.99	5.08	3.35	5.98	3.98	6.33	7.26	6.31	4.48	1.12
InternVL2	76B	3	5.57	6.95	5.11	5.81	7.37	5.96	3.61	3.83	6.48	4.66	6.05	6.05	6.30	4.23	0.93
MiniCPM	8B	6	5.42	7.18	5.37	5.46	6.23	4.46	5.35	3.34	4.83	3.69	5.99	7.35	6.25	4.97	1.09
Qwen-VL-Chat	9B	12	5.13	6.43	5.87	5.40	4.80	5.11	5.58	2.98	4.10	3.12	5.51	7.19	6.07	4.50	1.01
InternLM-XC2-VL	7B	7	4.97	6.34	4.70	5.28	5.06	4.69	5.03	3.08	4.49	3.29	5.00	7.21	5.92	4.56	0.88
DeepSeek-VL	7B	11	4.70	6.53	5.52	5.10	3.98	3.87	4.19	2.50	3.96	2.58	5.46	7.15	5.83	4.47	1.02
Monkey-Chat	9B	10	4.70	6.04	4.88	5.57	4.66	4.18	4.96	3.01	4.00	2.61	4.87	6.29	6.15	3.96	0.96
ShareGPT4V	13B	14	4.39	5.93	4.61	5.16	3.77	4.04	4.58	2.45	3.73	2.19	5.05	6.39	5.36	3.79	1.08
LLava-v1.5	13B	15	4.31	6.02	4.56	4.46	3.85	3.69	4.72	2.46	3.69	2.10	4.75	6.21	5.60	3.96	1.05
Yi-VL	34B	13	4.25	4.79	4.78	5.19	3.33	3.58	4.47	2.42	3.25	2.08	4.72	6.61	5.87	4.13	1.20
Phi-3-Vision	4B	9	4.08	4.48	3.53	4.75	4.10	3.48	3.16	2.56	4.40	2.85	4.34	5.51	5.85	4.07	0.86
InstructBLIP	9B	16	3.31	4.11	4.61	4.11	2.77	3.05	2.92	1.76	2.58	1.12	3.36	3.17	5.42	4.02	1.09
GPT-4o without image	-	-	2.13	1.11	1.57	1.22	1.73	1.53	1.17	1.29	2.88	1.14	1.99	3.50	5.14	3.41	-

Table 3: Evaluation results on AlignMMBench. For each column, the highest score is **bold**, while the second highest score is underlined. The "Ref." column indicates the relative ranking of these models on the `https://rank.opencompass.org.cn/leaderboard-multimodal`, dominated by primarily English benchmarks. Table 4 presents the detailed versions and architectures of these open-sourced models.

- **Robust Convergence**: These models exhibit a strong capacity for alignment, with both their visual and language components demonstrating robustness to achieve high levels of performance.
- **Component-Centric**: While the visual and language parts in these models exhibit notable strengths, further advancements are necessary to enhance their multimodal alignment.
- **Alignment-Driven**: These models display effective multimodal alignment; however, their overall performance remains constrained by inherent limitations in their visual or language parts.
- **Baseline Limited**: These models exhibit limited capabilities in both alignment and their visual/language parts, necessitating further improvements to enhance performance.

5.3.3 Analysis on category-level performance

Single-turn scenarios. Focusing on VLMs with average scores above 6.0, we observe that they perform well in tasks such as description, OCR, and writing. These tasks require VLMs to understand images but do not include complex reasoning or computation. However, they do not perform well in tasks involving reasoning and problem solving. For instance, GPT4o (OpenAI., 2024) scores only 4.44 in reasoning and 5.81 in problems, indicating frequent response errors. These observations suggest that while current top-tier VLMs can comprehend images and integrate information from images and texts, they struggle with test cases that demand complex reasoning and computation.

Multi-turn scenarios. In the coherence task, GPT-4o (OpenAI., 2024) demonstrates the best performance among all models. Most VLMs achieve scores above 6.0, indicating their ability to follow instructions from previous interactions and utilize information from the dialogue context effectively. However, all VLMs exhibit suboptimal performance in the incoherence task. This suggests that these models struggle to detect previous errors within the dialogue context and to make accurate corrections as guided by users.

Please see Appendix A.2 for more analysis.

6 Conclusion

In this paper, we introduce AlignMMBench, a comprehensive Chinese multimodal alignment benchmark comprising three high-level categories and thirteen subtasks. AlignMMBench includes 1,054 images and 4,978 question-answer pairs, encompassing both single-turn and multi-turn dialogue scenarios. To facilitate accurate and efficient evaluations, we developed a critique generation model, referred to as CritiqueVLM. Experimental results demonstrate that CritiqueVLM can assign scores aligned with human preferences and achieve superior performance compared to the widely used GPT-4. Additionally, we present a leaderboard featuring popular VLMs, highlighting potential directions for future improvements in VLMs. We anticipate that this dataset will further advance the development of multimodal language models.

Limitations

First, AlignMMBench functions as a alignment benchmark within Chinese context and does not evaluate the multilingual capabilities of VLMs. We plan to gather more images and questions in other languages to extend the scope of our benchmark. Second, due to the suboptimal performance of VLMs in scoring evaluation tasks (Chen et al., 2024a), we currently employ an LLM as our evaluator, which necessitates the use of reference answers that exclude image-derived information. In future work, we will explore the feasibility of using VLMs as reliable evaluators without requiring human-annotated reference answers.

References

Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Harkirat Behl, et al. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

01. AI, :, Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Heng Li, Jiangcheng Zhu, Jianqun Chen, Jing Chang, Kaidong Yu, Peng Liu, Qiang Liu, Shawn Yue, Senbin Yang, Shiming Yang, Tao Yu, Wen Xie, Wenhao Huang, Xiaohui Hu, Xiaoyi Ren, Xinyao Niu, Pengcheng Nie, Yuchi Xu, Yudong Liu, Yue Wang, Yuxuan Cai, Zhenyu Gu, Zhiyuan Liu, and Zonghong Dai. 2024. Yi: Open foundation models by 01.ai. *arXiv preprint arXiv:2403.04652*.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023a. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*.

Shuai Bai, Shusheng Yang, Jinze Bai, Peng Wang, Xingxuan Zhang, Junyang Lin, Xinggang Wang, Chang Zhou, and Jingren Zhou. 2023b. Touchstone: Evaluating vision-language models by language models. *arXiv preprint arXiv:2308.16890*.

Yonatan Bitton, Hritik Bansal, Jack Hessel, Rulin Shao, Wanrong Zhu, Anas Awadalla, Josh Gardner, Rohan Taori, and Ludwig Schmidt. 2023. Visit-bench: A benchmark for vision-language instruction following inspired by real-world use. *arXiv preprint arXiv:2308.06595*.

Dongping Chen, Ruoxi Chen, Shilin Zhang, Yinuo Liu, Yaochen Wang, Huichi Zhou, Qihui Zhang, Pan Zhou, Yao Wan, and Lichao Sun. 2024a. Mllm-as-a-judge: Assessing multimodal llm-as-a-judge with vision-language benchmark. *arXiv preprint arXiv:2402.04788*.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. 2024b. Are we on the right way for evaluating large vision-language models? *arXiv preprint arXiv:2403.20330*.

Lin Chen, Jisong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. 2023a. Sharegpt4v: Improving large multimodal models with better captions. *arXiv preprint arXiv:2311.12793*.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. 2024c. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*.

Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. 2024d. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *arXiv preprint arXiv:2404.16821*.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. 2023b. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. *arXiv preprint arXiv:2312.14238*.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. *arXiv preprint arXiv:2305.06500*.

Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Bin Wang, Linke Ouyang, Xilin Wei, Songyang Zhang, Haodong Duan, Maosong Cao, Wenwei Zhang, Yining Li, Hang Yan, Yang Gao, Xinyue Zhang, Wei Li, Jingwen Li, Kai Chen, Conghui He, Xingcheng Zhang, Yu Qiao, Dahua Lin, and Jiaqi Wang. 2024. Internlm-xcomposer2: Mastering free-form text-image composition and comprehension in vision-language large model. *arXiv preprint arXiv:2401.16420*.

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2022. Glm: General language model pretraining with autoregressive blank infilling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 320–335.

Paul Duncum. 2004. Visual culture isn't just visual: Multiliteracy, multimodality and meaning. *Studies in art education*, 45(3):252–264.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, and Rongrong Ji. 2024. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*.

Zhangwei Gao, Zhe Chen, Erfei Cui, Yiming Ren, Weiyun Wang, Jinguo Zhu, Hao Tian, Shenglong Ye, Junjun He, Xizhou Zhu, et al. 2024. Mini-internvl: A flexible-transfer pocket multimodal model with 5% parameters and 90% performance. *arXiv preprint arXiv:2410.16261*.

Wentao Ge, Shunian Chen, Guiming Hardy Chen, Junying Chen, Zhihong Chen, Nuo Chen, Wenya Xie, Shuo Yan, Chenghao Zhu, Ziyue Lin, Song Dingjie, Xidong Wang, Anningzhe Gao, Zhang Zhiyi, Jianquan Li, Xiang Wan, and Benyou Wang. 2024. Mllm-bench: Evaluating multimodal llms with per-sample criteria. *arXiv preprint arXiv:2311.13951*.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering.

Tuomo Hiippala, Malihe Alikhani, Jonas Haverinen, Timo Kalliokoski, Evanfiya Logacheva, Serafina Orekhova, Aino Tuomainen, Matthew Stone, and John A Bateman. 2021. Ai2d-rst: A multimodal corpus of 1000 primary school science diagrams. *Language Resources and Evaluation*.

Jinyi Hu, Yuan Yao, Chongyi Wang, Shan Wang, Yinxu Pan, Qianyu Chen, Tianyu Yu, Hanghao Wu, Yue Zhao, Haoye Zhang, Xu Han, Yankai Lin, Jiao Xue, Dahai Li, Zhiyuan Liu, and Maosong Sun. 2023. Large multilingual models pivot zero-shot multimodal learning across languages. *arXiv preprint arXiv:2308.12038*.

Chu-Ren Huang, Keh-Jiann Chen, Li-Li Chang, and Feng-Yi Chen. 1997. Segmentation standard for chinese natural language processing. In *International Journal of Computational Linguistics & Chinese Language Processing, Volume 2, Number 2, August 1997*, pages 47–62.

EunJeong Hwang and Vered Shwartz. 2023. Memecap: A dataset for captioning and interpreting memes. *arXiv preprint arXiv:2305.13703*.

Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. 2014. Referitgame: Referring to objects in photographs of natural scenes. In *Proc. of Empirical Methods in Natural Language Processing*.

Seungone Kim, Jamin Shin, Yejin Cho, Joel Jang, Shayne Longpre, Hwaran Lee, Sangdoo Yun, Seongjin Shin, Sungdong Kim, James Thorne, et al. 2023. Prometheus: Inducing fine-grained evaluation capability in language models. *Proc. of International Conference on Learning Representations*.

Seungone Kim, Juyoung Suk, Shayne Longpre, Bill Yuchen Lin, Jamin Shin, Sean Welleck, Graham Neubig, Moontae Lee, Kyungjae Lee, and Minjoon Seo. 2024. Prometheus 2: An open source language model specialized in evaluating other language models. *arXiv preprint arXiv:2405.01535*.

Zhang Li, Biao Yang, Qiang Liu, Zhiyin Ma, Shuo Zhang, Jingxu Yang, Yabo Sun, Yuliang Liu, and Xiang Bai. 2023. Monkey: Image resolution and text label are important things for large multi-modal models. *arXiv preprint arXiv:2311.06607*.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Proc. of European Conference on Computer Vision*.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024. Visual instruction tuning. *Proc. of Neural Information Processing Systems*.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. 2023a. Mmbench: Is your multi-modal model an all-around player? *arXiv preprint arXiv:2307.06281*.

Yuliang Liu, Zhang Li, Hongliang Li, Wenwen Yu, Mingxin Huang, Dezhi Peng, Mingyu Liu, Mingrui Chen, Chunyuan Li, Lianwen Jin, et al. 2023b. On the hidden mystery of ocr in large multimodal models. *arXiv preprint arXiv:2305.07895*.

Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Hao Yang, Yaofeng Sun, Chengqi Deng, Hanwei Xu, Zhenda Xie, and Chong Ruan. 2024a. Deepseek-vl: Towards real-world vision-language understanding. *arXiv preprint arXiv:2403.05525*.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. 2024b. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. *Proc. of International Conference on Learning Representations*.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. 2019. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proc. of Computer Vision and Pattern Recognition*.

Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. 2022. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. *arXiv preprint arXiv:2203.10244*.

Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. 2021. Docvqa: A dataset for vqa on document images.

OpenAI. 2024. Gpt4o. *https://openai.com/index/hello-gpt-4o*.

Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2015. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proc. of International Conference on Computer Vision*.

Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. 2020. Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 3505–3506.

Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. 2022. A-okvqa: A benchmark for visual question answering using world knowledge.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. 2019. Towards vqa models that can read.

Amanpreet Singh, Guan Pang, Mandy Toh, Jing Huang, Wojciech Galuba, and Tal Hassner. 2021. Textocr: Towards large-scale end-to-end reasoning for arbitrary-shaped scene text.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, et al. 2023. Cogvlm: Visual expert for pretrained language models. *arXiv preprint arXiv:2311.03079*.

Ruyi Xu, Yuan Yao, Zonghao Guo, Junbo Cui, Zanlin Ni, Chunjiang Ge, Tat-Seng Chua, Zhiyuan Liu, and Gao Huang. 2024. LLaVA-UHD: an lmm perceiving any aspect ratio and high-resolution images. *arXiv preprint arXiv:2403.11703*.

Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. 2016. Modeling context in referring expressions. In *Proc. of European Conference on Computer Vision*.

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

Tianyu Yu, Haoye Zhang, Yuan Yao, Yunkai Dang, Da Chen, Xiaoman Lu, Ganqu Cui, Taiwen He, Zhiyuan Liu, Tat-Seng Chua, and Maosong Sun. 2024. Rlaif-v: Aligning mllms through open-source ai feedback for super gpt-4v trustworthiness. *arXiv preprint arXiv:2405.17220*.

Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. 2023b. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. 2023. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. *arXiv preprint arXiv:2311.16502*.

Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, et al. 2022. Glm-130b: An open bilingual pre-trained model. *arXiv preprint arXiv:2210.02414*.

Xinlu Zhang, Yujie Lu, Weizhi Wang, An Yan, Jun Yan, Lianke Qin, Heng Wang, Xifeng Yan, William Yang Wang, and Linda Ruth Petzold. 2023. Gpt-4v(ision) as a generalist evaluator for vision-language tasks. *arXiv preprint arXiv:2311.01361*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2024. Judging llm-as-a-judge with mt-bench and chatbot arena. *Proc. of Neural Information Processing Systems*.

A Appendix

A.1 Detailed category definition

We provide a detailed definition of each task in AlignMMBench below, briefed in Figure 2.

Perception and understanding. This category involves answering questions by synthesizing information from images and world knowledge. It comprises six different tasks:

- **Description.** This task evaluates the model's ability to provide general descriptions of an image. All questions are constructed by sampling from a list of description requests.
- **Recognition.** This task requires the model to recognize objects in the image, covering a wide range of topics, including plants, animals, famous people, and films.
- **Counting.** This task requires the model to count specific objects in the image.
- **OCR.** This task assesses the model's ability to understand Chinese characters within images.

- **Meme.** Memes consist of widely shared images paired with text, using visual metaphors to convey meanings different from the original images (Hwang and Shwartz, 2023). This task aims to test whether models can truly understand the meaning of images, which is useful for examining the model's ability to grasp the profound metaphor between images and texts.
- **Knowledge.** This task tests the model's world knowledge. VLMs are required to provide comprehensive answers based on their internal knowledge, such as introducing the history of a tourist attraction or providing public information about a celebrity.

Reasoning and analysis. Beyond perception & understanding, this section focuses on assessing the model's capabilities in information processing and analysis. These tasks often require numerical calculations or logical reasoning to provide accurate responses. This category is divided into five distinct tasks:

- **Reasoning.** This task assesses the model's logical reasoning capacity, requiring it to make accurate choices and interpretations by identifying corresponding patterns in images, as illustrated in Figure 4.
- **Chart.** This task requires the model to parse information from charts in images, converting it to specified formats or calculating question-specific statistical characteristics.
- **Problem.** This task involves solving high school-level problems across various disciplines such as Mathematics, Physics, Chemistry, and Biology. To differentiate from the OCR category, we ensure that most of these problems cannot be correctly answered based solely on the text in the images.
- **Comparison.** This task requires models to identify differences between two sub-images or compare objects in different locations. Models need to locate objects and analyze the differences between them.
- **Writing.** Writing based on images requires comprehensive analysis and creative abilities. In this task, the model is required to compose poems, advertisements, speeches, and other forms of text based on the images.

Dialogue context. In real-world user scenarios, models are typically required to generate responses based on dialogue context. However, this remains an underexplored area in existing multimodal benchmarks. We broadly categorize this capability into two tasks:

- **Coherence.** In practical applications, users prefer that VLMs retain and utilize the entire dialogue context to generate coherent responses. This task focuses on evaluating the model's ability to follow instructions and incorporate information from dialogue history.
- **Incoherence.** When a model provides incorrect or irrelevant responses, users usually highlight these errors and expect the model to correct their replies. Therefore, this task assesses the model's self-correcting abilities.

A.2 Analysis on different backbone

To better illustrate the importance of our benchmark in Chinese, we conducted a survey of the open-source models featured in our leaderboard. The results are presented in Table 4. By integrating their underlying architectures with their performance metrics from the leaderboard, we make these observations:

- The training corpus plays an important role in model performance. As shown in the Table 4, models primarily trained in Chinese consistently outperform their English counterparts on AlignMMBench. Consequently, it can be inferred that previous evaluations of Chinese multimodal models based on English benchmarks may not have been sufficient and satisfactory.

- For both Chinese and English models, there is a generally positive correlation between model size and performance scores, indicating that larger and more recent models tend to achieve better results.

- Due to the variation in training corpora across different models, it is challenging to assess the specific impact of different model architectures. This issue represents an important research topic that extends beyond the scope of this paper's discussion.

A.3 Radar chart

To better analyze these VLMs' performance on different tasks, we draw an radar chart in Figure 5. Observations from this chart:

- GPT4o outperforms other models in most tasks, with CogVLM2 consistently ranking

Model	Size	Backbone/version	Main Language	Avg Scores
Qwen2-VL	72B	Qwen2-72B+QwenViT	Chinese	6.51
CogVLM2	19B	EVA2-CLIP-E + Adapter + MLP + LLaMA3 with vision Expert	Chinese	5.81
InternVL-V1.5	26B	ViT-L/14 + InternLM2	Chinese	5.62
InternVL2	76B	InternViT-6B + Llama-3-70B-Instruct	Chinese	5.57
MiniCPM-V2.5	8B	SigLIP + resampler + LLaMA3	Chinese	5.42
Qwen-VL-Chat	9B	Vit-BigG + CrossAttn + Qwen	Chinese	5.13
XComposer2V	7B	ViT-L/14 + InternLM2-7B	Chinese	4.97
DeepSeek-VL	7B	SAM-B & SigLIP-L + Adapter + DeepSeek	Chinese	4.70
Monkey-Chat	9B	Vit-BigG + resampler + Qwen	Chinese	4.70
ShareGPT 4V	13B	ViT-L/14 + MLP + LLaMA2	English	4.39
LLava1.5	13B	ViT-L/14 + MLP + LLaMA2	English	4.31
Yi-VL	34B	ViT-H/14 + MLP + Yi	Chinese	4.25
Phi3 Vision	4B	Vit-L/14 + Phi-3-mini	English	4.08
InstructBlip	9B	ViT-g/14 + QFormer + LLaMA-7B	English	3.31

Table 4: Summary table of different open-source model backbones.

second. Notably, in tasks such as "Problem" and "Meme", GPT4o establishes a significant lead over other VLMs.

- Existing VLMs could make good performance in "Perception & Understanding" category, but still need improvement on "Reasoning & Analysis" and "Dialogue Context".

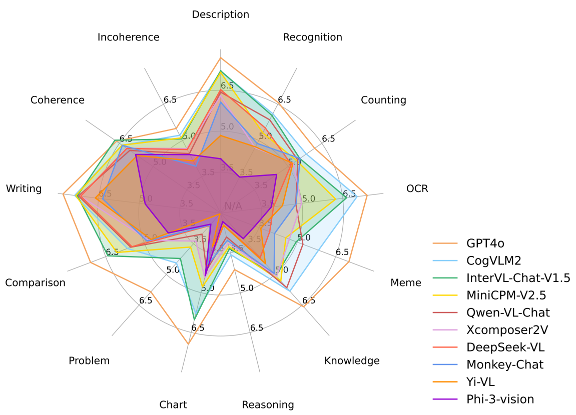


Figure 5: Radar chart of leaderboard results.

A.4 Prompt rewrite strategy

In this section, we provide a detailed implementation and example of our prompt rewriting strategy. For test cases in the "Perception & Understanding" and "Reasoning & Analysis" categories, we employ GLM4 to generate multiple different prompts while preserving the original intent of the seed prompt. Subsequently, we manually review all prompts to ensure they meet our requirements. The prompt used in our rewriting strategy is illustrated in Figure 8.

Figure 9 illustrates an example of equivalent prompts for the same images, where the model's scores varied significantly. This demonstrates that our rewrite strategy enhances the robustness of our alignment evaluation, effectively minimizing the fluctuations in results attributed to the model's inherent randomness. This improvement ensures more consistent and reliable assessments across different tests.

A.5 Analysis on question and reference answer lengths

In this section, we present the distribution of length for questions and reference answers in our Align-MMBench dataset. Results are illustrated in Figure 6.

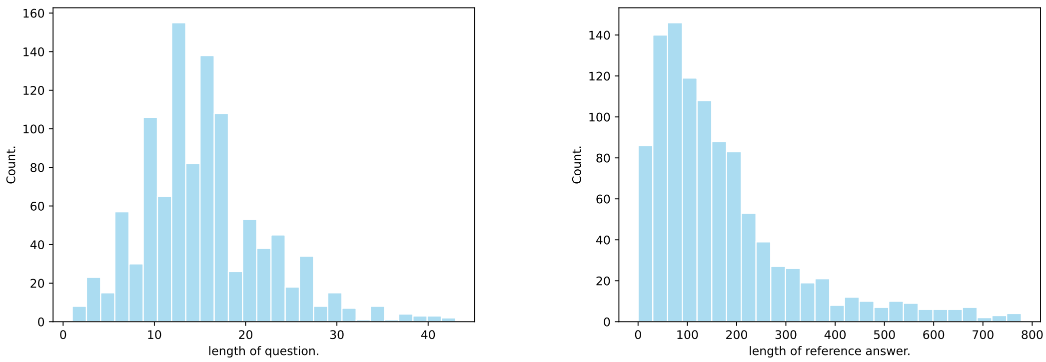
A.6 More details of data generation

A.6.1 Image collection

We conducted this process in the following steps: 1) Researchers manually defined 13 task types, and construct keywords for each tasks. 2) Web crawlers collected over 20k images using these keywords and removed duplicates based on their hash values. 3) Researchers manually selected images varying in style and content.

Below are several examples of the keywords in step 1:

- Description: Poster, Scenery, Flowchart, Map, Cartoon.
- Knowledge: Movie, Celebrity, Software icon, Famous attraction, Book.
- Chart: Flowchart, Subway map, Bar graph, Table, Index status, Data summary table.
- Recognition: Fruit, Flower, Plant, Animal, Food, Simple drawing.



(a) Distribution of questions.

(b) Distribution of reference answers.

Figure 6: Length distribution of questions and reference answers.

- OCR: Handwritten, Documents, Signs, Billboards, Math formulas.
- Counting: Fruit stall, Collections, Crowds.
- Reasoning: Puzzles, IQ test, Reasoning, Brain teaser.
- Problem: Math exercises, Reading comprehension problems, Mechanical problems, Biology problems.
- Comparison: Car comparison, Home products comparison, Product comparison, Cosmetics comparison.
- Writing: Scenery, Countryside, Forest, Urban, Wildlife.
- Meme: Meme.

A.6.2 Human annotation

Guidelines The guidelines provided to human annotators are fundamentally consistent with the prompt detailed in A.8. Furthermore, annotators were supplied with typical characteristics and illustrative examples corresponding to different scoring levels.

Participants and workload The annotators participating in out work are either employees or interns at Zhipu AI. Approximately ten individuals are involved in the annotation process, and each annotator processes about 60 entries per day. To ensure the high quality of the annotation results, additional personnel perform quality checks on the annotations at the end of each day.

A.7 Examples

A.7.1 Image examples

In this section, we randomly selecte multiple images to demonstrate the diversity of visual tasks, as depicted in Figure 7. These images illustrate that our AlignMMBench includes a variety of examples that emphasize unique features of the Chinese context. These include: 1) images related to Optical Character Recognition (OCR) and 2) the identification and description of Chinese celebrities, tourist attractions, and holidays. This suggests that our AlignMMBench serves as an excellent supplement to the predominantly English benchmarks currently available.

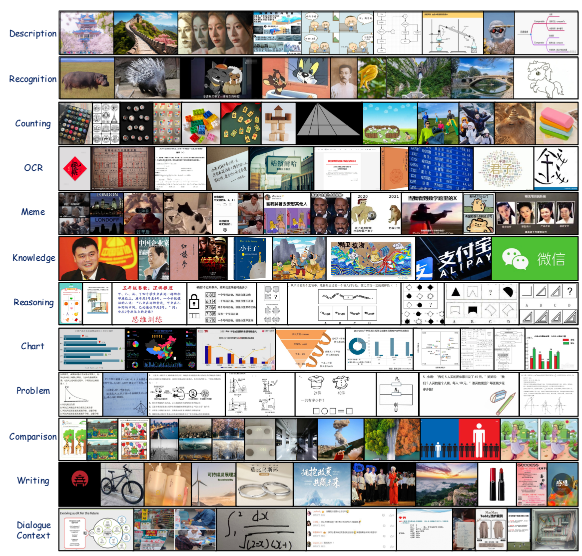


Figure 7: Image examples of different tasks.

你是一个精通汉语的语言专家。现在我会给你一个用户关于某张图片内容的提问，您需要在不改变提问的整体含义的前提下对该提问进行改写。也就是说，若某个人看到改写前和改写后的提问，他的回答应该是一致的。
[提问内容开始]{question}[提问内容结束]
您需要给出5 次不同的改写结果。输出格式要求输出5 行，每行包含其中的一个改写结果。请注意不要输出任何其他的字符。

You are a language expert proficient in Chinese. I will now provide you with a user's query about the content of a specific picture. You need to rewrite the query without altering its overall meaning. That is, if someone were to read both the original and the rewritten queries, their responses should be consistent.

[Question begin]
{seed question}
[Question end]

You need to provide five different rewrites of the query. The output format should be five lines, each containing one of the rewrites. Please ensure no other characters are output.

Figure 8: The prompt used in prompt-rewrite strategy.

Figure 9: An example of equivalent prompts for the same image.

A.7.2 Evaluation examples

To more effectively illustrate the evaluation capabilities of CritiqueVLM in comparision to GPT-4, we have chosen an evaluation example in Figure 10.

A.8 Evaluation prompt

Our evaluation prompt consists of general instructions and task-specific prompts, which are listed as follows.

General instruction is illustrated in Figure 11. This prompt encompasses the scoring range, criteria, general hints, and output format. We employ an in-context example to enable the critic model to infer the content of the image from the question and reference answers. Furthermore, it encourages the model to thoughtfully consider details not found in the reference answers rather than immediately deeming them incorrect.

Task-specific instruction is illustrated in Figure 12. We manually identify the errors made by the API model and construct specific hints to guide it towards accurate evaluations.

Figure 10: A specific comparative example between CritiqueVLM and GPT4 evaluation.

问答数据

{对话历史信息}

- 用户提问：{问题}
- 问题类型：{问题类型}

[参考答案开始]{参考答案}[参考答案结束]

[AI助手回答开始]{AI助手回答}[AI助手回答结束]

目标

您是一个擅长评价问答质量的专家。请您以客观公正的评判者身份，结合用户提问和参考答案内容，评估AI助手回答的质量。

指令

- 评分要求：

1. 您的评估应考虑到正确性、相关性和详细级别等因素。开始您的评估时，请将AI助手回答与参考答案进行比较，并从中找出AI助手回答可能存在的不合理内容，在整体评分结束后提供简短的解释说明。
2. 对于每次评分请求，您需要进行1到10的评分，评分规则如下：整体上，模型回答的质量越高，则分数越高。当AI助手回答完全没有遵循用户的文本指令内容，答非所问、自言自语，总分必须是1分；当AI助手回答和用户提问相关，但内容有害（如：不友好，违反安全性）或语句不通顺不流畅（如：存在大量重复，语言整体混乱），则总分必须是2分；当AI助手回答和用户提问相关，只是回答完全错误，则总分只能是3分。当AI助手回答和用户提问相关，回答有一半以上信息存在事实性错误，则总分应该在4～ 5分之间。当AI助手回答和用户提问相关，只是有一半以下的信息存在事实性错误，或回答无事实性错误但只满足了用户提问的部分要求，则总分应该是6分。当AI助手回答和用户提问相关，只有极个别错误或回答正确但过于简略，总分应当在7～ 8分之间。只有当AI助手回答完成了用户提问的所有要求，且提到的所有信息均正确，只是遗漏了一些并不重要的信息，且这些信息不影响回答内容的优劣判断下，总分可以给到9～ 10分。
3. 判断回答与问题是否相关或者回答是否正确时，应该充分考虑语言的同义性，等价性和近似语义。
4. 整体上，在回答结果正确的条件下，有合理分析的回答得分应该更高。

- 注意事项：

由于用户提问属于"{问题类型}"类型，因此，在进行打分和解释理由时，应该充分考虑以下注意事项：

{与类别相关的规则}

特别地：由于您无法查看图片内容，因此没有提供图片给您。但AI 助手在回答问题时，是根据图片内容进行回答的。因此当AI助手回答中存在参考答案中没有提到的图片相关信息时，您不能直接认为它是错误的，而是需要结合上述注意事项以及您现有的知识进行综合评价。比如：

- 问题：这幅卡通图片里的鱼是已经煮熟了吗？我们如何判断？
- 参考答案：是的，这个卡通图中的鱼已经煮熟了，因为它是不透明的，用叉子很容易剥离，并且内部温度已经达到145华氏度。
- AI助手：是的，这幅图片中的鱼已经煮熟了。我们可以通过鱼的颜色、质地和温度来判断。鱼的颜色是粉红色，质地变得松散并且容易用叉子夹碎，温度达到了145°F，这些都是鱼已经煮熟的迹象。
- 评分&理由：9分。首先，AI助手的结论与参考答案一致。其次，通过颜色和温度判断鱼是否熟了符合生活经验。整体上，AI助手的回答结论正确，逻辑成立，可以给一个较高的分数。

输出格式

您必须按照以下JSON 格式输出回答：{{ "Rating": , "Reason":}}

除了JSON内容外，请不要输出任何其他字符。并且，应该使用中文描述Reason 部分。

Q&A Data

{dialogue history}

- User Question: {question}
- Question Type: {questiontype}

[Reference Begin]{reference answer}[Reference End]

[AI Assistant Answer Begin]{AI Assistant Answer}[AI Assistant Answer End]

Objects

You are an expert at evaluating the quality of Q&A responses. As an objective and fair judge, evaluate the AI assistant's response based on the user's question and the reference answer.

Instructions

- Scoring Requirements:

1. Your evaluation should consider accuracy, relevance, and the level of detail. Begin your assessment by comparing the AI assistant's answer to the reference answer, identify any unreasonable content in the AI assistant's response, and provide a brief explanation after completing the overall score.
2. For each scoring request, you need to score from 1 to 10, with the rules as follows: The higher the quality of the model's response, the higher the score. If the AI assistant completely ignores the user's text instructions, answers unrelated questions or rambles, the total score must be 1; if the AI assistant's response is relevant to the user's question but contains harmful content (such as unfriendly, violating safety) or the sentence is not smooth (such as a lot of repetition, overall language chaos), then the total score must be 2; if the AI assistant's response is relevant to the user's question but completely incorrect, then the total score can only be 3. If the AI assistant's answer is relevant to the user's question and more than half of the information contains factual errors, the total score should be between 4 and 5. If the AI assistant's response is relevant to the user's question, but less than half of the information contains factual errors, or the response has no factual errors but only partially meets the user's question, then the total score should be 6. If the AI assistant's response is relevant to the user's question, has only a few minor errors, or is correct but too brief, the total score should be between 7 and 8. Only when the AI assistant's response meets all the requirements of the user's question, and all the mentioned information is correct, but some unimportant information is omitted and does not affect the judgment of the response quality, the total score can be 9 to 10.
3. When judging whether the answer is relevant to the question or whether the answer is correct, the synonymy, equivalence, and approximate semantics of the language should be fully considered.
4. Overall, if the answer is correct, a well-reasoned response should score higher.

- Points to Note:

Since the user's question falls under the "{questiontype}" category, when scoring and explaining reasons, you should consider the following points:

{Task-Specific Rules}

Especially: Since you cannot view image content, no images are provided for you. However, the AI assistant's answers are based on the content of the pictures. Therefore, when there is information related to the picture in the AI assistant's answer that is not mentioned in the reference answer, you cannot directly assume it is wrong but need to combine the above points and your existing knowledge for a comprehensive evaluation. For example:

- Question: Is the fish in this cartoon picture cooked? How can we tell?
- Reference Answer: Yes, the fish in this cartoon picture is cooked because it is opaque, easily peels off with a fork, and the internal temperature has reached 145 degrees Fahrenheit.
- AI Assistant: Yes, the fish in this picture is cooked. We can tell by the color, texture, and temperature of the fish. The fish's color is pink, the texture has become loose and easy to shred with a fork, and the temperature has reached 145°F, which are all signs that the fish is cooked.
- Score & Reason: 9. Firstly, the AI assistant's conclusion is consistent with the reference answer. Secondly, judging whether the fish is cooked based on its color and temperature aligns with real-life experiences. Overall, the AI assistant's response is correct and logically sound, warranting a high score.

Output Format

You must output the response in the following JSON format: {{"Rating": , "Reason":}}

Apart from the JSON content, please do not output any other characters. And, the Reason part should be described in Chinese.

Figure 11: The general prompt template for evaluation.

描述类: 1. 在评分时，应该充分考虑AI助手回答的组织条理性，逻辑性，语言流畅性和内容完整性。当AI助手回答内容存在不完整时，可以根据不完整程度进行酌情减分，但不能直接认为是AI助手回答是错误的。2. 由于参考答案是对某张图片内容的完整或者部分描述，因此参考答案可能存在描述不全的情况。当AI助手回答中存在参考答案中不存在的内容时，可以适当怀疑其内容的合理性，但不能直接认为新增内容是错误的。
推理类: 1. AI助手回答此类问题时应该提供合理的解释，尤其是问题要求给出理由时。2. 在评分时，应该首先判断AI助手回答的结论是否正确，若结论错误，可以判定其回答错误，此时应直接给一个低分；若结论正确，再结合其解释的合理性与逻辑性进行综合评分。
识别类: 1. 这类问题的回答重点在于识别结果的正确与否，且用户提问和参考答案内容均默认围绕图片进行。您应该耐心地从AI助手回答中找到针对问题的关键答案。当AI助手回答中结果与参考答案语义一致时，务必给高分甚至满分。2. 即使AI助手的回答和参考答案相比有多余的内容，只要AI助手的识别结果正确，回答中存在和参考答案语义一致的部分，且其余部分符合逻辑，就应当给高分甚至满分。3. 若回答中对识别结果有合理的描述或者推测，能够酌情加分。当然，不能超过评分规定中的10分。4. 识别文本内容时，除非题目中特别强调，否则不应该将翻译后的文字视为错误结果。5. 对于数字，应该注意等价转换，比如0.1 = 10%
计数类: 1. 这类问题的回答重点在于计数结果的正确与否，且用户提问和参考答案内容均默认围绕图片进行。您应该耐心地从AI助手回答中找到针对问题的关键答案。当AI助手回答中结果与参考答案一致时，务必给高分甚至满分。反之，结果只要不同时，不论差距有多小都必须视为完全错误，必须给低分。2. 即使AI助手的回答和参考答案相比有很多多余的内容，只要AI助手回答中存在和参考答案语义一致的部分，且其余部分符合逻辑，就应当给高分甚至满分。3. 若回答中对识别结果有不合理的描述或者推测，应该酌情减分。
图表类: 1. 由于您无法查看图片，所以请务必将AI助手回答与参考答案进行比较分析。2. 对于格式转换题，首先关注回答是否符合新格式要求，其次关注回答的内容是否正确。3. 对于数字，应该注意等价转换，比如0.1 = 10%
对比类: 1. 若题目是要求进行对比分析，则有一定组织格式的回答质量优于无组织格式的回答。
创作类: 1. 如果提问是要求根据图片写一个故事，那么即便AI助手的回答和参考答案差异较大，也不应该直接给1～4分。而应该根据回答中故事本身的流畅性、戏剧性、有趣程度和与用户提问的关联度等进行打分。
智力类: 1. 请注意检查用户提问和AI助手回答的一致性，如果AI助手答非所问，直接给低分。2. 如果提问类似“这道题怎么做？”，“图片中的题怎么做？”说明用户希望AI助手解决图片中的问题。此时请仔细对比参考答案和AI助手回复来判断AI助手是否真正解决了问题。
梗图理解: 1. 这类问题的重点在于评测AI助手能否正确理解该梗图的有趣点。所以你需要仔细比对AI助手和参考答案对图片的解释和理解是否是一种意思，如果是则要打高分；2. 如果AI助手完全没有解释图片为何有趣，或解释的内容和参考回答相比不足以让用户体会到这张图片背后的真实含义，应该给低分。
世界知识: 1. 这类问题的重点在于评测AI助手是否拥有图片以外的世界知识，所以当AI助手回答了一些参考答案中没有的内容时，你不能直接认为它是错误的，而是要结合逻辑是否通顺，与用户提问是否相关，和你自己拥有的知识等方面综合评判。
OCR理解: 1. 如果用户提问要求提取或识别内容，则此时应该严格要求答案与参考答案相同，此时不需要考虑“AI助手回答中存在参考答案中没有提到的图片相关信息”，此时只要参考答案与AI助手回答有差距，直接给一个低分。2. 如果用户提问要求提取或识别内容，此时你只需要检验AI助手回答和参考答案是否匹配，并不需要对回答本身是否包含错误进行检查。3. 注意判断AI助手回答与参考答案是否语义是一致的，只是语种。若是回答跟参考答案是不同语种的相同表达，则不应该以此理由给1～4分。
多轮对话: 1. "对话历史"中呈现了用户和AI助手之间的历史会话内容。2. 这里问题的重点在于评测AI助手能否在多轮对话中利用之前对话的知识，遵循先前对话的指令。所以您需要深刻理解"对话历史"的每一轮对话，将其与当前的"用户提问"和答案进行对比，给出综合的得分。3. 如果多轮对话历史中用户要求AI助手修复自己回答的错误，需要仔细观察本次AI助手回复是否能意识到自己的错误并给出正确的答案，如果没有意识到自己的错误或只是承认错误但未给出修正过的答案，则只能获得低分。

Description: 1. In grading, the organization, logic, fluency, and completeness of the AI assistant's responses should be fully considered. Points may be deducted for incomplete responses, but incompleteness should not be equated with incorrectness. 2. Since the reference answer may describe an image either partially or completely, it may be incomplete. If the AI assistant includes content not found in the reference answer, its accuracy should be scrutinized, but additional content should not be automatically deemed incorrect.
Reasoning: 1. The AI assistant should provide reasonable explanations, especially when asked to justify a response. 2. When grading, first assess if the conclusion of the AI assistant's response is correct. If incorrect, the response should be considered wrong and receive a low score. If correct, further evaluate the reasonableness and logic of the explanation.
Recognition: 1. These questions focus on the accuracy of identification results, with both user questions and reference answers usually centered around an image. Carefully identify key responses from the AI. If the AI's results align semantically with the reference answer, award high or full marks. 2. Even if the AI's response includes extra information compared to the reference answer, as long as the identification is correct and parts of the response align semantically with the reference, it deserves high or full marks. 3. Reasonable descriptions or conjectures about identification results in the response can merit additional points, but must not exceed the maximum allowed. 3. When identifying text content, translations should not be considered incorrect unless specifically stated in the question. 5. With numbers, equivalent conversions should be noted, e.g., 0.1 equals 10%.
Counting: 1. The focus here is on the accuracy of counting results. Patience is required to extract key responses from the AI. If the AI's results match the reference answer, award high or full marks. If not, even slight discrepancies must be treated as completely incorrect, warranting low scores. 2. Even with extraneous content in the AI's response, as long as part of it semantically matches the reference answer and the rest is logical, award high or full marks. 3. Unreasonable descriptions or conjectures about counting results should lead to deductions.
Chart: 1. As you cannot view images, always compare and analyze the AI assistant's responses against the reference answers. 2. For format conversion questions, prioritize whether the response meets the new format requirements and then check if the response content is correct. 3. For numerical data, ensure equivalent conversions, e.g., 0.1 equals 10%.
Comparison: 1. If the question involves comparative analysis, organized responses should be rated higher than those without structure.
Writing: 1. If asked to create a story based on an image, significant differences from the reference answer should not automatically result in scores between 1 and 4. Instead, grade based on the narrative flow, drama, interest, and relevance to the query. Intellectual Type:
Problem: 1. Consistency between the user's query and the AI's response should be verified. Irrelevant answers should receive low scores. 2. For questions like 'How to solve this problem?' or 'How to address the problem shown in the image?', it's crucial to compare the AI's solution to the reference answer to determine if the problem was effectively addressed.
Meme: 1. These questions assess whether the AI understands the humor of a meme. Carefully compare the AI's interpretation with the reference answer. If they align and the AI captures the essence of the humor, award high marks. If the AI fails to explain why the image is humorous or if its explanations fall short of conveying the underlying meaning compared to the reference answer, award low scores.
Knowledge: 1. When the AI provides information not in the reference answer, do not automatically dismiss it as incorrect. Instead, evaluate its relevance to the query, logical coherence, and consistency with known facts.
OCR: 1. When tasked with text extraction or recognition, the response must match the reference answer exactly. If there is a discrepancy, regardless of additional context provided, score low. 2. For text extraction or recognition questions, only assess whether the AI's response matches the reference answer. Do not penalize for potential errors unless the semantic content differs.
Dialogue Context: 1. "Dialogue history" shows the interaction history between the user and the AI. Assess whether the AI uses knowledge from previous exchanges to adhere to the user's ongoing directives. 2. If a user asks the AI to correct an error in its previous responses, observe whether the AI acknowledges and corrects the error in its new response. If it fails to do so, award a low score.

Figure 12: Task-specific rules in evaluation prompt.