

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

**Yuhang Wu<sup>1,\*†</sup>, Wenmeng Yu<sup>2,\*</sup>, Yean Cheng<sup>3,†</sup>, Yan Wang<sup>2</sup>,**  
**Xiaohan Zhang<sup>2</sup>, Jiazheng Xu<sup>1,†</sup>, Ming Ding<sup>2</sup>, Yuxiao Dong<sup>1</sup>**

<sup>1</sup>Tsinghua University    <sup>2</sup>Zhipu AI    <sup>3</sup>Peking University

## 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, a comprehensive alignment benchmark specifically designed for emerging Chinese VLMs. This benchmark is meticulously curated from real-world scenarios and Chinese 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. 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 <https://alignmmbench.github.io/>.

## 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 [2, 35, 4]. 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, researchers in the Chinese community have made significant advancements and proposed multiple impressive VLMs, such as QwenVL [4], CogVLM [35], MiniCPM [16], Intern-VL [10], Intern-XComposer2V [11], Deepseek-VL [24], and Yi-VL [3]. To facilitate objective comparisons and evaluations among these models, researchers mainly utilize the following benchmarks: 1) general evaluation sets such as MME [13] and MMBench [22], and 2) domain-specific evaluation sets such as MathVista [25] and MMMU [41]. However, these benchmarks predominantly assess model capabilities through non-verbal methods and lack detailed evaluations of alignment performance, particularly in Chinese contexts. These limitations hinder a comprehensive understanding and optimization of emerging Chinese VLMs.

\*Equal contribution.

†Done as intern at Zhipu AI.

\*\*Corresponding author [yuxiaod@tsinghua.edu.cn](mailto:yuxiaod@tsinghua.edu.cn)

<sup>1</sup>MathVista contains 6.57% non-English questions, such as Chinese and Persian.

Table 1: Comparisons between AlignMMBench and other benchmarks.

Benchmark	Dataset					Evaluation	
	Size	Language	Multi Category	Dialogue Context	Open Ended	Judge Method	Metric
Ai2D [15]	4,903	English	✗	✗	✓	multi-choice	Accuracy
LLaVABench [21]	150	English	✓	✗	✓	GPT	Score
MathVista [25]	6,141	English <sup>1</sup>	✗	✗	✓	GPT	Accuracy
MME [13]	2,800	English	✓	✗	✗	yes/no	Accuracy
MMBench [22]	3,217	Eng. & Chi.	✓	✗	✗	multi-choice	Accuracy
MMMU [41]	11,500	English	✗	✗	✗	multi-choice	Accuracy
MMStar [8]	1,500	English	✓	✗	✗	multi-choice	Accuracy
MM-Vet [40]	205	English	✓	✗	✓	GPT	Score
OCRBench [23]	1,000	English	✗	✗	✓	string match	Accuracy
TouchStone [5]	908	English	✓	✗	✓	GPT	Score
VisIT-Bench [6]	592	English	✓	✗	✓	GPT	Elo / Win
<b>AlignMMBench</b>	<b>4,978</b>	<b>Chinese</b>	<b>✓</b>	<b>✓</b>	<b>✓</b>	<b>CritiqueVLM</b>	<b>Score</b>

Constructing a high-quality Chinese multimodal alignment benchmark is fully challenging. First, unlike the English context, which benefits from traditional benchmarks such as VQAv2 [14], TextVQA [33], and ChartQA [27], clean and publicly available Chinese multimodal corpora are exceedingly scarce. Therefore, it is necessary to collect and construct these resources from scratch. Second, the ambiguity inherent in the Chinese language is greater, requiring multiple annotators to repeatedly verify one single high-quality Chinese multimodal corpus. Nevertheless, building a Chinese multimodal benchmark is essential and promising. The characteristics of images and the underlying world knowledge can vary significantly between different languages. Consequently, relying solely on English datasets makes it difficult to comprehensively evaluate Chinese VLMs.

To address this gap, we introduce AlignMMBench, a comprehensive evaluation benchmark specifically designed to assess the alignment capabilities of Chinese VLMs. AlignMMBench is meticulously curated from real-world scenarios and Chinese internet resources, encompassing thirteen specific tasks across three high-level categories, including both single-turn and multi-turn dialogue scenarios. Considering the diversity of query forms in real-world user scenarios, we introduce a LLM-based synonym expansion strategy, which transforms a single query into multiple stylistically different but semantically equivalent questions. As illustrated in Figure 1, AlignMMBench ultimately comprises 1,054 images and 4,978 question-answer pairs.

As the absence of fixed 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 [2], to score the model responses. However, the API-based GPT-4 operates as a closed-source black box, resulting in limited control for developers. Consequently, when the underlying model of the API undergoes changes, the corresponding scoring results are also affected. Considering these limitations, we fine-tuned an evaluator based on the open-source ChatGLM3-6B, named CritiqueVLM. Help with the detailed prompt rules and chain-of-thought reasoning, CritiqueVLM achieves higher consistency with human scoring compared to GPT-4.

Based on AlignMMBench and CritiqueVLM, we evaluate ten popular VLMs that support Chinese. Our findings are as follows: 1) Most of VLMs can easily follow task instructions; 2) These VLMs excel in perception and understanding but perform poorly in reasoning and analysis; 3) Within dialogue contexts, VLMs exhibit suboptimal performance in the incoherence task compared to the coherence task, indicating that VLMs struggle to detect previous errors.

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 public first alignment benchmark specifically designed for Chinese VLMs.

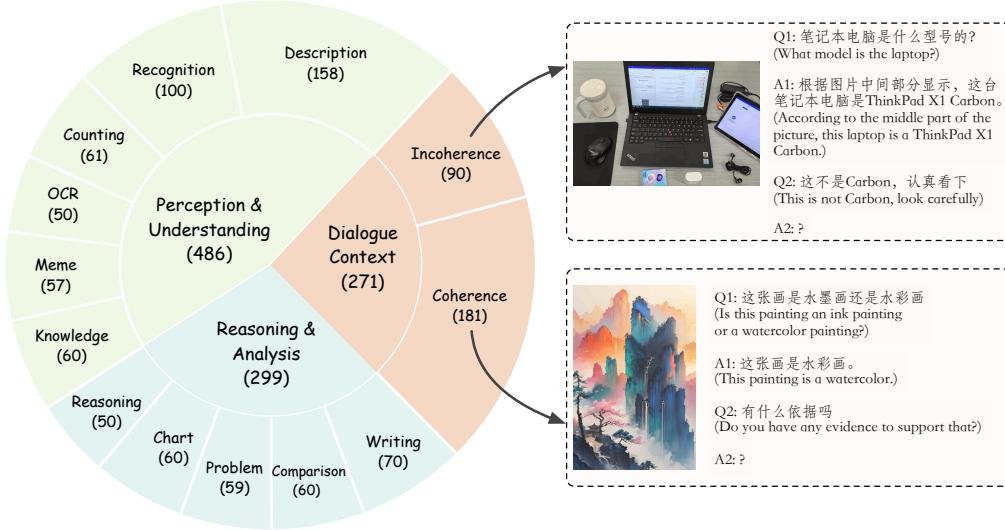


Figure 1: Categories and examples of AlignMMBench. The chart on the left displays the categories of AlignMMBench, encompassing three main categories and thirteen specific tasks. The numbers below each category or task indicate the total number of images included. The right side of the pie chart presents two examples, illustrating instances from the incoherence and coherence categories.

- To improve the controllability of evaluation scores, we introduce the CritiqueVLM , a ChatGLM3-6B based evaluator that has been rule-calibrated and carefully finetuned. With human judgements, its evaluation consistency surpasses that of GPT-4.
- We benchmark representative VLMs on AlignMMBench . Beyond their performance on AlignMMBench , 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 [20, 30], Visual Grounding [18, 37], Visual Question Answering [32, 26, 28] and Optical Character Recognition [34]. The emergence of general VLMs underscores the necessity for comprehensive benchmarks capable of evaluating integrated capabilities.

To address these needs, recent works strive to propose various benchmarks containing complicated questions that expect VLMs to generate accurate and high-quality responses. These benchmarks can be categorized into two types based on their questions, including domain-specific and general scenarios. 1) Domain-specific benchmarks concentrate on particular subfields. For example, discipline reasoning [41, 25], Optical Character Recognition (OCR) [23], chart understanding [27, 15], and the hallucination test [21]. 2) General benchmarks provide multi-dimensional and diverse assessments, encompassing MME [13], MMBench [22], MMStar [8], MMVet [40], LLaVABench [8], VisIT-Bench [6], and TouchStone [5]. From another perspective, based on differences in result evaluation methods, the aforementioned benchmarks can be divided into deterministic and open-ended types. Deterministic benchmarks require the model to produce responses in a specific format through rule-based prompts. Common formats include multiple choice, yes / no, and strict matching. In contrast, open-ended benchmarks allow the model to generate unconstrained responses, which are then evaluated using advanced language models. 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.

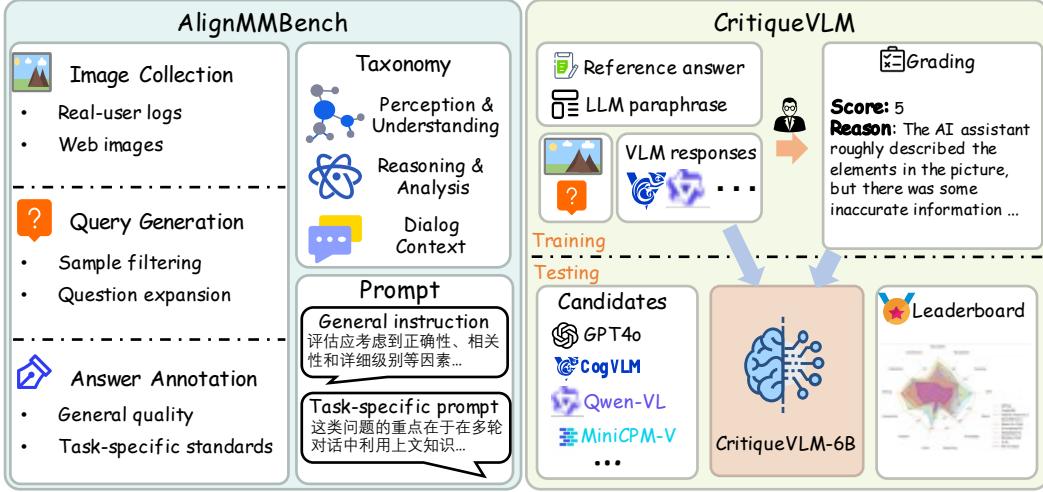


Figure 2: Overall framework of our work. **AlignMMBench**: Images are collected from real-world requests and Chinese websites. Following manual taxonomy and LLM-based question expansion, annotators generate reference answers within task-specific rules. **CritiqueVLM**: Rule-based scoring criteria are carefully designed, and then we train the CritiqueVLM on human-annotated 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 [25] utilizes an LLM to extract answers from responses for more precise comparisons with reference answers, while other works [21, 40, 5, 6] adopt the LLM-as-a-Judge [43] approach. This method prompts the LLM to act as a judge, scoring responses or selecting the better response between two candidates. Existing VLM benchmarks use LLM-as-a-Judge methods with a unified evaluation prompt across all tasks and leverage GPTs [2] as their evaluator.

However, evaluating numerous models within a large benchmark can be costly and inefficient, with results potentially becoming unstable due to API-based model updates. 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 [12] 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 1.

**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 provide the exact number of specific objects in the image.
- **OCR.** This task assesses the model’s ability to recognize and 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 [17]. 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 evaluates the model’s logical reasoning abilities in visual contexts. We collect visual reasoning questions that require the model to make correct choices and interpretation by identifying corresponding patterns in images, such as Case 2 in Figure 3.
- **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 its replies. This task assesses the model’s self-correcting abilities when users indicate its mistakes.

### 3.2 Dataset construction

As illustrated in Figure 2, the construction of AlignMMBench involves three primary steps: image collection, query generation, and answer annotation. This section provides a concise overview of these steps.

**Image collection.** To establish a diverse and high-quality benchmark, we collect images through two primary methods: web image crawlers and real-world user requests. Each image underwent rigorous manual screening to ensure suitability for public use and to avoid inclusion in 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 [12] to rephrase the seed questions without altering their original purpose. This method was applied specifically to single-turn questions.

### Case 1



详细解析一下这张图像。  
(Analyze this image in detail.)

这幅画是一幅名画，描绘了一位身穿黑色衣服的女士，她的头发是黑色的，眼睛是棕色的。她双手交叉放在胸前，背景是一片山脉和河流。这幅画的风格看起来很古典，使用了柔和的色彩和细腻的笔触。  
(This painting is a famous painting that depicts a lady in black clothes, with black hair and brown eyes. She has her hands crossed on her chest, with a mountain range and river in the background. The style of this painting looks very classical, using soft colors and delicate brushstrokes.)



Score: 5

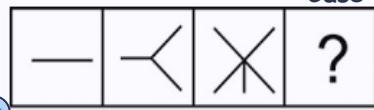
**Reason:** AI助手的回答与用户提问的内容相关，它描述了图像的基本特征，如女士的穿着、背景山脉和河流。然而，与参考答案相比，AI助手的回答在详细程度和准确性方面存在不足。AI助手没有指出图像中的错觉趣味元素，如将女性的脖子和前胸看作小猫、手看作兔子和头发装饰柱体的形象。此外，AI助手没有提到从拍摄者视角看，图片左侧背景山体中似乎能看到一张脸的细节。因此，虽然AI助手的回答满足了问题的大致要求，但在内容的完整性和详细程度方面有所欠缺，故给予中等水平的评分。

(The AI assistant's response is relevant to the user's question, describing the basic features of the image, such as the lady's attire, the background mountains, and the river. However, compared to the reference answer, the AI assistant's response lacks detail and accuracy. The AI assistant did not point out the illusionary fun elements in the image, such as interpreting the woman's neck and chest as a kitten, her hand as a rabbit, and her hair as decorative columns. Additionally, the AI assistant did not mention the detail that from the photographer's perspective, it seems there is a face visible in the left background mountain of the image. Therefore, although the AI assistant's response broadly meets the requirements of the question, it lacks completeness and detail in the content, hence it is given a moderate rating.)

### Case 2

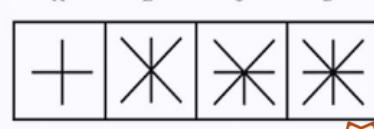


根据图片中的规律，问号部分应选择哪个选项?  
(According to the pattern in the image, which option should be chosen for the question mark section?)



考虑到第一行和第二行图案的相似性和差异性，第一行的最后一个图案是由两个相反方向的线段组成的，而第二行的最后一个图案是由四个相互交叉的线段组成的。因此，问号部分应该选择D选项。

(Considering the similarity and difference between the patterns in the first and second rows, the last pattern in the first row consists of two line segments in opposite directions, while the last pattern in the second row is made up of four intersecting line segments. Therefore, the question mark section should select option D.)



Score: 3

**Reason:** AI助手在回答问题时，虽然考虑了图片中图案的相似性和差异性，但其推理过程与参考答案给出的规律不符。参考答案中明确指出图形的射线数依次为1、3、5，都是奇数，并且根据这一规律推断问号处图形的射线数应为7。而AI助手没有提供这样的逻辑推理，而是根据图案的组成方式进行了选择，这与问题要求的推理规律不一致，因此评分较低。同时，由于评分者无法查看图片内容，无法验证AI助手回答的准确性，但根据回答的规律和逻辑性，可以判断其回答的相关性不高。

(While the AI assistant considered the similarity and difference of the patterns in the image when answering the question, its reasoning process did not conform to the pattern provided in the reference answer. The reference answer clearly stated that the number of rays in the shapes is 1, 3, 5, all odd numbers, and inferred that the number of rays in the shape at the question mark should be 7 based on this pattern. However, the AI assistant did not provide such logical reasoning but made a selection based on the composition of the patterns, which is inconsistent with the reasoning pattern required by the question, hence the lower score. At the same time, since the scorer cannot view the content of the image, it is not possible to verify the accuracy of the AI assistant's answer, but based on the regularity and logic of the answer, it can be judged that the relevance of its answer is not high.)

Figure 3: Critical examples from CritiqueVLM.

**Answer annotation.** We adopt a two-phase annotation process. First, human annotators craft responses based on the corresponding images. Second, 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.

Table 2: Mean absolute error ( $e$ ), Pearson ( $r$ ), Spearman ( $\rho$ ), Kendall ( $\tau$ ) correlations, Fuzzy division scores ( $f$ ) and Strict division scores ( $s$ ) are calculated on the internal and external test sets. The best performance among the methods is **bold**.

Testcase	Internal							External						
	Metric	$e$	$r$	$\rho$	$\tau$	$f$	$s$	$e$	$r$	$\rho$	$\tau$	$f$	$s$	
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*		<b>0.818</b>	<b>0.846</b>	<b>0.838</b>	<b>0.740</b>	<b>0.747</b>	<b>0.646</b>	<b>1.146</b>	<b>0.778</b>	<b>0.782</b>	<b>0.671</b>	<b>0.670</b>	<b>0.511</b>	

## 4 CritiqueVLM

In this section, we develop a LLM-based evaluator, referred to as CritiqueVLM , to automatically assess responses based on the questions and detailed reference answers. In addition to assigning a score between 1 and 10, CritiqueVLM also provides a chain-of-thought explanation, shown in Figure 3. CritiqueVLM is fine-tuned from ChatGLM3-6B, an open-source model in the ChatGLM3 series [12, 42].

### 4.1 Training data construction

**Prompt design.** Following [43], we devise a prompt system consisting of general 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 our supplementary materials.

**Scoring data collection.** We collect the training corpus based on CogVLM [35] and QwenVL-Chat [4]. 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 [12, 42]. 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 [2] and annotated scores provided by human experts.

### 4.2 Supervised fine-tuning

We select ChatGLM3-6B [12, 42] as the base model and fine-tune it on the training dataset. The DeepSpeed [31] 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) [2], as baselines of CritiqueVLM , which are widely used in other benchmarks. Additionally, given AlignMMBench is a Chinese benchmark, **ChatGLM3-6B** [12, 42] is also chosen as a baseline.

## 5.2 Main results

In our experiment, we employ six statistical metrics to assess the agreement between human annotated scores and model generated scores. Common four statistical metrics include the mean absolute error ( $e$ ), Pearson ( $r$ ), Spearman ( $\rho$ ), and Kendall correlation coefficient ( $\tau$ ). 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 3, demonstrating that CritiqueVLM achieves superior performance with only 6 billion parameters. Additionally, Figure 3 provides two examples of scoring by CritiqueVLM , validating its ability to generate reasonable critical scores and detailed explanations.

## 6 Leaderboard

We benchmark a range of popular VLMs, including GPT4o [29], CogVLM2 [35], InternVL-Chat-V1.5 [10, 9], MiniCPM-V2.5 [38, 16, 36, 39], Qwen-VL-Chat [4], XComposer2V [11], DeepSeek-VL [24], Monkey-Chat [19], Yi-VL [3], and Phi-3-Vision [1]. Results are shown in Table 3.

### 6.1 Analysis on overall result

The average scores of VLMs in Table 3 range from 3.9 to 6.6. According to our scoring criteria, we observe that most VLMs can understand question requirements and generate responses related to the images. Scores below 5.0 indicate numerous errors in the VLMs' responses. Besides, GPT4o [29] achieves the best performance across most tasks, and CogVLM2 [35] secures the second-best performance. Overall, these scores disparity between GPT4o and other models suggests that GPT4o's capabilities significantly surpass those of the other models.

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.

### 6.2 Analysis on category-level performance

**Single-turn scenarios.** Focusing on VLMs with average scores above 5.0, we observe that they perform well in tasks such as description, OCR, comparison, and writing. These tasks require VLMs to understand images but do not necessitate complex reasoning or computation. However, they do not perform well in tasks involving Reasoning and Problem-solving. For instance, GPT4o [29] scores only 4.13 in reasoning and 5.85 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, InternVL-Chat-V1.5 [10, 9] 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.

## 7 Conclusion

In this paper, we introduce AlignMMBench , a comprehensive Chinese multimodal alignment benchmark comprising three high-level categories and thirteen subtasks. AlignMMBench includes

Table 3: Evaluation results on AlignMMBench. Scores in each category are the average scores of test cases. 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.

Models	Ref.	Avg	Perception & Understanding					Reasoning & Analysis					Context		
			Des.	Rec.	Cou.	OCR.	Mem.	Kno.	Rea.	Cha.	Pro.	Com.	Wri.	Coh.	Inc.
GPT-4o [29]	1	<b>6.57</b>	<b>7.69</b>	<b>6.56</b>	<b>6.18</b>	<b>7.4</b>	<b>7.02</b>	<b>6.58</b>	<b>4.13</b>	<b>6.94</b>	<b>5.85</b>	<b>7.09</b>	<b>7.81</b>	<b>6.63</b>	<b>5.5</b>
CogVLM2 [35]	5	<u>5.86</u>	7.19	<u>6.11</u>	<u>5.81</u>	<u>7.03</u>	<u>5.83</u>	<u>5.82</u>	3.57	5.79	4.44	5.65	7.38	6.38	<u>5.22</u>
InternVL-V1.5 [10]	2	5.73	<u>7.21</u>	6.04	5.52	6.64	4.96	5.06	3.34	<u>6.02</u>	4.22	<u>6.41</u>	7.28	<b>6.70</b>	5.07
MiniCPM-V2.5 [16]	3	5.43	7.15	5.32	5.38	6.24	4.55	5.32	3.27	4.78	3.66	6.06	7.34	6.38	5.11
Qwen-VL-Chat [4]	8	5.13	6.43	5.86	5.43	4.77	5.22	5.65	2.90	4.15	3.03	5.52	7.24	6.04	4.46
XComposer2V [11]	4	4.98	6.29	4.71	5.23	4.98	4.69	5.09	3.14	4.43	3.37	4.81	7.17	6.19	4.67
DeepSeek-VL [24]	7	4.75	6.52	5.52	5.14	3.95	3.92	4.18	2.46	3.98	2.61	5.47	7.16	6.17	4.63
Monkey-Chat [19]	6	4.71	6.06	4.88	5.53	4.78	4.11	4.95	3.03	3.94	2.58	4.88	6.35	6.17	3.94
Yi-VL [3]	9	4.24	4.84	4.74	5.21	3.28	3.57	4.47	2.46	3.27	2.04	4.72	6.62	5.71	4.16
Phi-3-Vision [1]	-	3.90	3.99	3.49	4.49	3.87	3.34	3.25	2.32	4.36	2.54	4.05	4.78	5.78	4.42

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 ten popular VLMs supporting Chinese, highlighting potential directions for future improvements in VLMs. We anticipate that this dataset will further advance the development of multimodal language models.

## 8 Limitations

First, AlignMMBench functions as a benchmark for Chinese language 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 [7], 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.

## Acknowledgement

We extend our sincere gratitude to the data annotation team at Zhipu AI for their invaluable assistance and support. All GPU compute and API expenses are generously covered by Zhipu AI.

## References

- [1] M. Abdin, S. A. Jacobs, A. A. Awan, J. Aneja, A. Awadallah, H. Awadalla, N. Bach, A. Bahree, A. Bakhtiari, H. Behl, et al. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*, 2024.
- [2] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.
- [3] . AI, :, A. Young, B. Chen, C. Li, C. Huang, G. Zhang, G. Zhang, H. Li, J. Zhu, J. Chen, J. Chang, K. Yu, P. Liu, Q. Liu, S. Yue, S. Yang, S. Yang, T. Yu, W. Xie, W. Huang, X. Hu, X. Ren, X. Niu, P. Nie, Y. Xu, Y. Liu, Y. Wang, Y. Cai, Z. Gu, Z. Liu, and Z. Dai. Yi: Open foundation models by 01.ai. *arXiv preprint arXiv:2403.04652*, 2024.
- [4] J. Bai, S. Bai, S. Yang, S. Wang, S. Tan, P. Wang, J. Lin, C. Zhou, and J. Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 2023.

- [5] S. Bai, S. Yang, J. Bai, P. Wang, X. Zhang, J. Lin, X. Wang, C. Zhou, and J. Zhou. Touchstone: Evaluating vision-language models by language models. *arXiv preprint arXiv:2308.16890*, 2023.
- [6] Y. Bitton, H. Bansal, J. Hessel, R. Shao, W. Zhu, A. Awadalla, J. Gardner, R. Taori, and L. Schmidt. Visit-bench: A benchmark for vision-language instruction following inspired by real-world use. *arXiv preprint arXiv:2308.06595*, 2023.
- [7] D. Chen, R. Chen, S. Zhang, Y. Liu, Y. Wang, H. Zhou, Q. Zhang, P. Zhou, Y. Wan, and L. Sun. Mllm-as-a-judge: Assessing multimodal llm-as-a-judge with vision-language benchmark. *arXiv preprint arXiv:2402.04788*, 2024.
- [8] L. Chen, J. Li, X. Dong, P. Zhang, Y. Zang, Z. Chen, H. Duan, J. Wang, Y. Qiao, D. Lin, et al. Are we on the right way for evaluating large vision-language models? *arXiv preprint arXiv:2403.20330*, 2024.
- [9] Z. Chen, W. Wang, H. Tian, S. Ye, Z. Gao, E. Cui, W. Tong, K. Hu, J. Luo, Z. Ma, et al. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *arXiv preprint arXiv:2404.16821*, 2024.
- [10] Z. Chen, J. Wu, W. Wang, W. Su, G. Chen, S. Xing, M. Zhong, Q. Zhang, X. Zhu, L. Lu, B. Li, P. Luo, T. Lu, Y. Qiao, and J. Dai. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. *arXiv preprint arXiv:2312.14238*, 2023.
- [11] X. Dong, P. Zhang, Y. Zang, Y. Cao, B. Wang, L. Ouyang, X. Wei, S. Zhang, H. Duan, M. Cao, W. Zhang, Y. Li, H. Yan, Y. Gao, X. Zhang, W. Li, J. Li, K. Chen, C. He, X. Zhang, Y. Qiao, D. Lin, and J. Wang. Internlm-xcomposer2: Mastering free-form text-image composition and comprehension in vision-language large model. *arXiv preprint arXiv:2401.16420*, 2024.
- [12] Z. Du, Y. Qian, X. Liu, M. Ding, J. Qiu, Z. Yang, and J. Tang. 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, 2022.
- [13] C. Fu, P. Chen, Y. Shen, Y. Qin, M. Zhang, X. Lin, J. Yang, X. Zheng, K. Li, X. Sun, Y. Wu, and R. Ji. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*, 2024.
- [14] Y. Goyal, T. Khot, D. Summers-Stay, D. Batra, and D. Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. 2017.
- [15] T. Hiippala, M. Alikhani, J. Haverinen, T. Kalliokoski, E. Logacheva, S. Orekhova, A. Tuomainen, M. Stone, and J. A. Bateman. Ai2d-rst: A multimodal corpus of 1000 primary school science diagrams. *Language Resources and Evaluation*, 2021.
- [16] J. Hu, Y. Yao, C. Wang, S. Wang, Y. Pan, Q. Chen, T. Yu, H. Wu, Y. Zhao, H. Zhang, X. Han, Y. Lin, J. Xue, D. Li, Z. Liu, and M. Sun. Large multilingual models pivot zero-shot multimodal learning across languages. *arXiv preprint arXiv:2308.12038*, 2023.
- [17] E. Hwang and V. Schwartz. Memecap: A dataset for captioning and interpreting memes. *arXiv preprint arXiv:2305.13703*, 2023.
- [18] S. Kazemzadeh, V. Ordonez, M. Matten, and T. Berg. Referitgame: Referring to objects in photographs of natural scenes. In *Proc. of Empirical Methods in Natural Language Processing*, 2014.
- [19] Z. Li, B. Yang, Q. Liu, Z. Ma, S. Zhang, J. Yang, Y. Sun, Y. Liu, and X. Bai. Monkey: Image resolution and text label are important things for large multi-modal models. *arXiv preprint arXiv:2311.06607*, 2023.
- [20] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *Proc. of European Conference on Computer Vision*, 2014.

- [21] H. Liu, C. Li, Q. Wu, and Y. J. Lee. Visual instruction tuning. *Proc. of Neural Information Processing Systems*, 2024.
- [22] Y. Liu, H. Duan, Y. Zhang, B. Li, S. Zhang, W. Zhao, Y. Yuan, J. Wang, C. He, Z. Liu, et al. Mmbench: Is your multi-modal model an all-around player? *arXiv preprint arXiv:2307.06281*, 2023.
- [23] Y. Liu, Z. Li, H. Li, W. Yu, M. Huang, D. Peng, M. Liu, M. Chen, C. Li, L. Jin, et al. On the hidden mystery of ocr in large multimodal models. *arXiv preprint arXiv:2305.07895*, 2023.
- [24] H. Lu, W. Liu, B. Zhang, B. Wang, K. Dong, B. Liu, J. Sun, T. Ren, Z. Li, H. Yang, Y. Sun, C. Deng, H. Xu, Z. Xie, and C. Ruan. Deepseek-vl: Towards real-world vision-language understanding. *arXiv preprint arXiv:2403.05525*, 2024.
- [25] P. Lu, H. Bansal, T. Xia, J. Liu, C. Li, H. Hajishirzi, H. Cheng, K.-W. Chang, M. Galley, and J. Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. *Proc. of International Conference on Learning Representations*, 2024.
- [26] K. Marino, M. Rastegari, A. Farhadi, and R. Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proc. of Computer Vision and Pattern Recognition*, 2019.
- [27] A. Masry, D. X. Long, J. Q. Tan, S. Joty, and E. Hoque. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. *arXiv preprint arXiv:2203.10244*, 2022.
- [28] M. Mathew, D. Karatzas, and C. Jawahar. Docvqa: A dataset for vqa on document images. 2021.
- [29] OpenAI. Gpt4o. <https://openai.com/index/hello-gpt-4o>, 2024.
- [30] B. A. Plummer, L. Wang, C. M. Cervantes, J. C. Caicedo, J. Hockenmaier, and S. Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proc. of International Conference on Computer Vision*, 2015.
- [31] J. Rasley, S. Rajbhandari, O. Ruwase, and Y. He. 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, 2020.
- [32] D. Schwenk, A. Khandelwal, C. Clark, K. Marino, and R. Mottaghi. A-okvqa: A benchmark for visual question answering using world knowledge. 2022.
- [33] A. Singh, V. Natarajan, M. Shah, Y. Jiang, X. Chen, D. Batra, D. Parikh, and M. Rohrbach. Towards vqa models that can read. 2019.
- [34] A. Singh, G. Pang, M. Toh, J. Huang, W. Galuba, and T. Hassner. Textocr: Towards large-scale end-to-end reasoning for arbitrary-shaped scene text. 2021.
- [35] W. Wang, Q. Lv, W. Yu, W. Hong, J. Qi, Y. Wang, J. Ji, Z. Yang, L. Zhao, X. Song, et al. Cogvlm: Visual expert for pretrained language models. *arXiv preprint arXiv:2311.03079*, 2023.
- [36] R. Xu, Y. Yao, Z. Guo, J. Cui, Z. Ni, C. Ge, T.-S. Chua, Z. Liu, and G. Huang. LLaVA-UHD: an lmm perceiving any aspect ratio and high-resolution images. *arXiv preprint arXiv:2403.11703*, 2024.
- [37] L. Yu, P. Poirson, S. Yang, A. C. Berg, and T. L. Berg. Modeling context in referring expressions. In *Proc. of European Conference on Computer Vision*, 2016.
- [38] T. Yu, Y. Yao, H. Zhang, T. He, Y. Han, G. Cui, J. Hu, Z. Liu, H.-T. Zheng, M. Sun, et al. Rlhf-v: Towards trustworthy mllms via behavior alignment from fine-grained correctional human feedback. *arXiv preprint arXiv:2312.00849*, 2023.

- [39] T. Yu, H. Zhang, Y. Yao, Y. Dang, D. Chen, X. Lu, G. Cui, T. He, Z. Liu, T.-S. Chua, and M. Sun. Rlaif-v: Aligning mllms through open-source ai feedback for super gpt-4v trustworthiness. *arXiv preprint arXiv:2405.17220*, 2024.
- [40] W. Yu, Z. Yang, L. Li, J. Wang, K. Lin, Z. Liu, X. Wang, and L. Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*, 2023.
- [41] X. Yue, Y. Ni, K. Zhang, T. Zheng, R. Liu, G. Zhang, S. Stevens, D. Jiang, W. Ren, Y. Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. *arXiv preprint arXiv:2311.16502*, 2023.
- [42] A. Zeng, X. Liu, Z. Du, Z. Wang, H. Lai, M. Ding, Z. Yang, Y. Xu, W. Zheng, X. Xia, et al. Glm-130b: An open bilingual pre-trained model. *arXiv preprint arXiv:2210.02414*, 2022.
- [43] L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Proc. of Neural Information Processing Systems*, 2024.

## Appendix

### A Datasheet

#### A.1 Motivation

- **For what purpose was the dataset created?** Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.  
The dataset was created to evaluate the alignment ability of VLMs in Chinese. As far as we know, this is the public first alignment benchmark specifically designed for Chinese VLMs.
- **Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?**  
The datasets were created by ZhipuAI company.
- **Who funded the creation of the dataset?** If there is an associated grant, please provide the name of the grantor and the grant name and number.  
No grant.
- **Any other comments?**  
No.

#### A.2 Composition

- **What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.**  
The dataset comprises images paired with question-answer pairs. We divide them into 13 different tasks: Description, Recognition, Counting, OCR, Meme, Knowledge, Reasoning, Chart, Problem, Comparison, Writing, Coherence, and Incoherence.
- **How many instances are there in total (of each type, if appropriate)?**  
There are 1,054 images and 4,978 question-answer pairs. Detailed distribution in each task type is in our Introduction Section.
- **Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set?** If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable).  
The dataset is a sample from all the Chinese images and user logs.
- **What data does each instance consist of?** “Raw” data (e.g., unprocessed text or images) or features? In either case, please provide a description.  
Each instance in the dataset consists of an image, a question, and a reference answer.
- **Is there a label or target associated with each instance?** If so, please provide a description.  
Yes, each image-question pair is associated with a reference response that serves as the target for this instance.
- **Is any information missing from individual instances?** If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.  
No.
- **Are relationships between individual instances made explicit (e.g., users’ movie ratings, social network links)?** If so, please describe how these relationships are made explicit.  
N/A.
- **Are there recommended data splits (e.g., training, development/validation, testing)?** If so, please provide a description of these splits, explaining the rationale behind them.  
No. The entire AlignMMBench dataset is used to test VLMs’ ability, so there are no recommended data splits.

- *Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.*

Yes, the reference answers are annotated by humans and may contain variations due to the diversity of human expression.

- *Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (i.e., including the external resources as they existed at the time the dataset was created); c) are there any restrictions (e.g., licenses, fees) associated with any of the external resources that might apply to a dataset consumer? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.*

The dataset is self-contained. We gather images from the web and real-user logs, and generate questions and responses through human annotation.

- *Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor–patient confidentiality, data that includes the content of individuals’ non-public communications)? If so, please provide a description.*  
No.

- *Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.*  
No.

- *Any other comments?*

No.

### A.3 Collection Process

- *How was the data associated with each instance acquired? Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If the data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.*

The data associated with each instance are annotated by researchers and professional annotators at Zhipu AI.

- *What mechanisms or procedures were used to collect the data (e.g., hardware apparatus(es) or sensor(s), manual human curation, software program(s), software API(s))? How were these mechanisms or procedures validated?*

We employ web crawlers to gather images from the web, utilize the GLM4 API to expand the question set, and let annotators at Zhipu AI annotate reference answers.

- *If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?*

We manually select images to ensure high-quality images paired with question-answer sets.

- *Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?*

Image collection and question construction are carried out by researchers. And Zhipu AI's professional annotators annotate reference answers. We ensure that annotators are paid more than the minimum wage in our country.

- *Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.*

N/A.

- *Were any ethical review processes conducted (e.g., by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.*

No.

- *Any other comments?*

No.

#### A.4 Preprocessing/cleaning/labeling

- *Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remaining questions in this section.*

Yes. Each image underwent rigorous manual screening to ensure suitability for public use and to avoid inclusion in known SFT datasets or prior benchmarks. Researchers review preliminary annotations, and filter the low-quality annotations.

- *Was the “raw” data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)? If so, please provide a link or other access point to the “raw” data.*

No.

- *Is the software that was used to preprocess/clean/label the data available? If so, please provide a link or other access point.*

No.

- *Any other comments?*

No.

#### A.5 Uses

- *Has the dataset been used for any tasks already? If so, please provide a description.*

Not beyond this paper.

- *Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point.*

Yes. We will update it on <https://alignnmbench.github.io/>.

- *What (other) tasks could the dataset be used for?*

The dataset can be applied to other image-related tasks such as visual grounding, image understanding, and visual question answering.

- *Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a dataset consumer might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other risks or harms (e.g., legal risks, financial harms)? If so, please provide a description. Is there anything a dataset consumer could do to mitigate these risks or harms?*

No.

- *Are there tasks for which the dataset should not be used? If so, please provide a description.*

No.

- *Any other comments?*

No.

#### A.6 Distribution

- *Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.*

Yes, the dataset will be published later.

- *How will the dataset be distributed (e.g., tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?*

The dataset will be made available in JSONL format on HuggingFace. As of now, it does not have a DOI.

- *When will the dataset be distributed?*

The dataset will be distributed in July 2024.

- *Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing*

*terms or ToU, as well as any fees associated with these restrictions.*

Yes, under CC BY-NC-SA 4.0 license.

- **Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.**

No.

- **Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.**

No.

- **Any other comments?**

No.

## A.7 Maintenance

- **Who will be supporting/hosting/maintaining the dataset?**

The dataset will be hosted on HuggingFace.

- **How can the owner/curator/manager of the dataset be contacted (e.g., email address)?**

Email address:

- **Is there an erratum? If so, please provide a link or other access point.**

No.

- **Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to dataset consumers (e.g., mailing list, GitHub)?**

Yes, the dataset will be updated when errors in responses are reported. In the future, we may also add new instances or remove existing ones to ensure the dataset's accuracy and relevance.

- **If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were the individuals in question told that their data would be retained for a fixed period of time and then deleted)? If so, please describe these limits and explain how they will be enforced.**

N/A.

- **Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to dataset consumers.**

Yes, we will use HuggingFace to maintain version history.

- **If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to dataset consumers? If so, please provide a description.**

No.

- **Any other comments?**

No.

## B Radar chart

To better analyze these VLMs' performance on different tasks, we draw an radar chart in Figure 4. Observations from this chart:

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

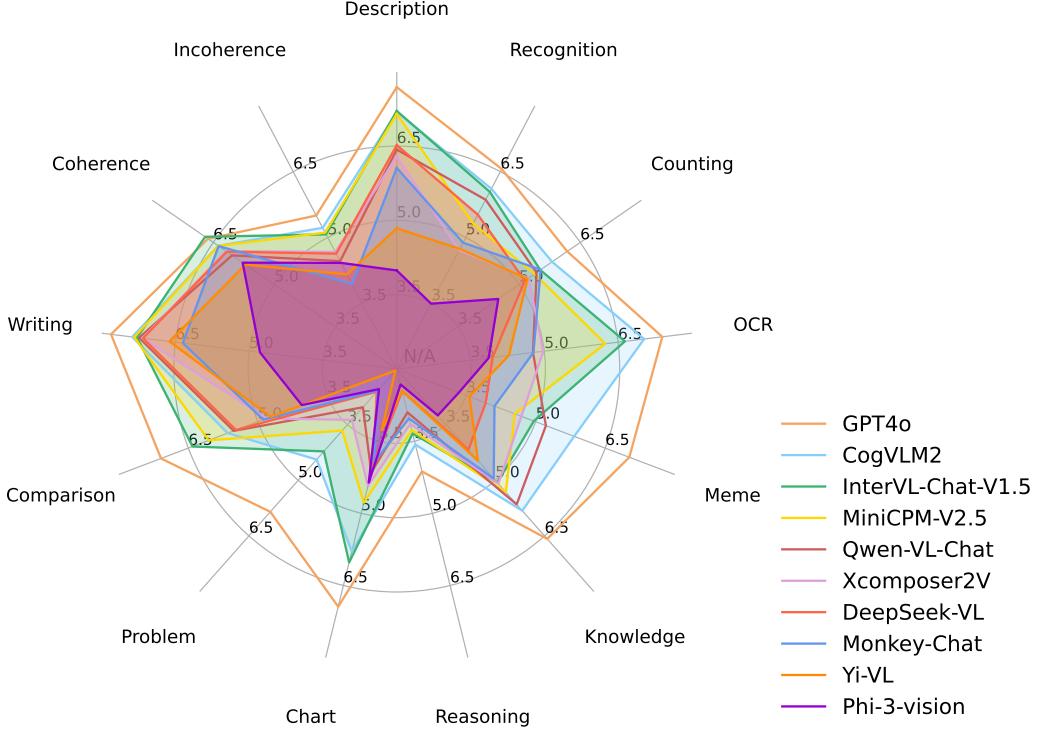


Figure 4: Radar chart of leaderboard results.

## C Image examples

In this section, we randomly select multiple images to illustrate the vision diversity of each task, shown as Figure 5.

## D 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 6.

Figure 7 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.

## E Evaluation prompt

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

**General instruction** is illustrated in Figure 8. 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.



Figure 5: 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 6: The prompt used in prompt-rewrite strategy.

**Task-specific instruction** is illustrated in Figure 9. We manually identify the errors made by the API model and construct specific hints to guide it towards accurate evaluations.

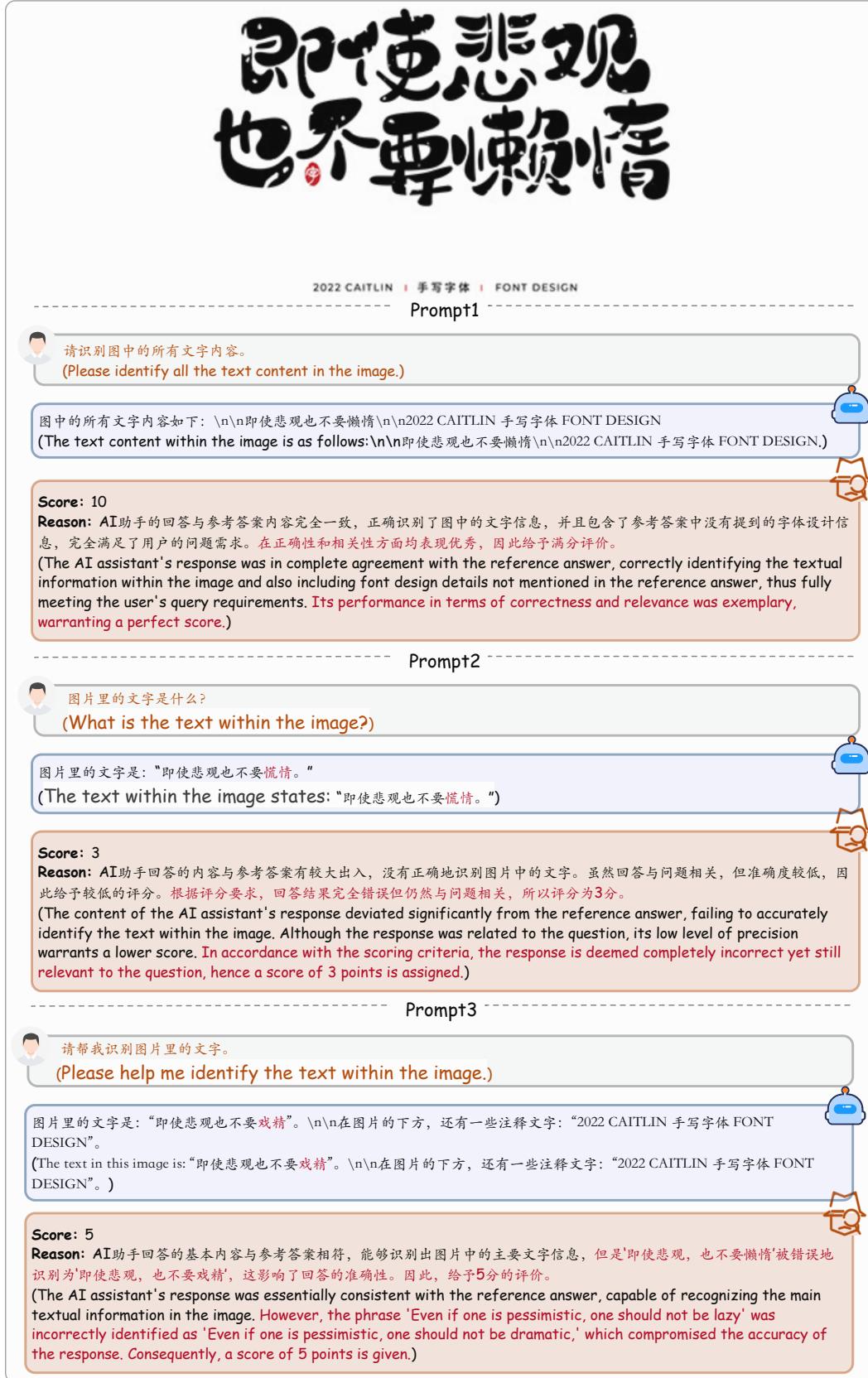


Figure 7: An example of equivalent prompts for the same image.

```

## 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，只是有一半以上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. 判断回答与问题是相关还是回答是否正确时，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。我们可以从鱼的颜色、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 8: 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很多多余的内容，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，不能超过评分规定的10分。4. 判别文本内容时，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，不论差距有多少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，那么即使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助手回答来判断AI助手是否真正解决了问题。

**根图理解:** 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。所以你需要仔细对比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助手回答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 9: Task-specific rules in evaluation prompt.