arXiv:2307.06281v5 [cs.CV] 20 Aug 2024

MMBench: Is Your Multi-modal Model an All-around Player?

Yuan Liu[1,*], **Haodong Duan**[1,*,‡], **Yuanhan Zhang**[2,*], **Bo Li**[2,*], **Songyang Zhang**[1,*],
Wangbo Zhao[4], **Yike Yuan**[5], **Jiaqi Wang**[1], **Conghui He**[1], **Ziwei Liu**[2,†], **Kai Chen**[1,†]
Dahua Lin[1,3,†]
[1]Shanghai AI Laboratory [2]Nanyang Technological University
[3] The Chinese University of Hong Kong [4] National University of Singapore
[5] Zhejiang University
[*] Equal Contribution [‡] Project Lead [†] Corresponding Author

Abstract

Large vision-language models (VLMs) have recently achieved remarkable progress, exhibiting impressive multimodal perception and reasoning abilities. However, effectively evaluating these large VLMs remains a major challenge, hindering future development in this domain. Traditional benchmarks like VQAv2 or COCO Caption provide quantitative performance measurements but lack fine-grained ability assessment and robust evaluation metrics. Meanwhile, subjective benchmarks, such as OwlEval, offer comprehensive evaluations of a model's abilities by incorporating human labor, which is not scalable and may display significant bias. In response to these challenges, we propose MMBench, a bilingual benchmark for assessing the multi-modal capabilities of VLMs. MMBench methodically develops a comprehensive evaluation pipeline, primarily comprised of the following key features: 1. MMBench is meticulously curated with well-designed quality control schemes, surpassing existing similar benchmarks in terms of the number and variety of evaluation questions and abilities; 2. MMBench introduces a rigorous CircularEval strategy and incorporates large language models to convert free-form predictions into pre-defined choices, which helps to yield accurate evaluation results for models with limited instruction-following capabilities. 3. MMBench incorporates multiple-choice questions in both English and Chinese versions, enabling an apples-to-apples comparison of VLMs' performance under a bilingual context. To summarize, MMBench is a systematically designed **objective** benchmark for a **robust** and **holistic** evaluation of vision-language models. We hope MMBench will assist the research community in better evaluating their models and facilitate future progress in this area. The evalutation code of MMBench has been integrated into VLMEvalKit [14]. [1]

1 Introduction

Recently, notable progress has been achieved within the realm of large language models (LLMs). For instance, the latest LLMs, such as OpenAI's ChatGPT and GPT-4 [37], have demonstrated remarkable reasoning capabilities that are comparable to, and in some cases, even surpass human capabilities. Drawing inspiration from these promising advancements in LLMs, large vision-language models (LVLMs) have also experienced a revolutionary transformation. Notable works, such as

---

[1] This is a revised version released in April 2024. It describes MMBench v1.1, a refined version of the MMBench (with better data quality). Please refer to https://arxiv.org/pdf/2307.06281v3 for the previous version, which is released in August 2023.

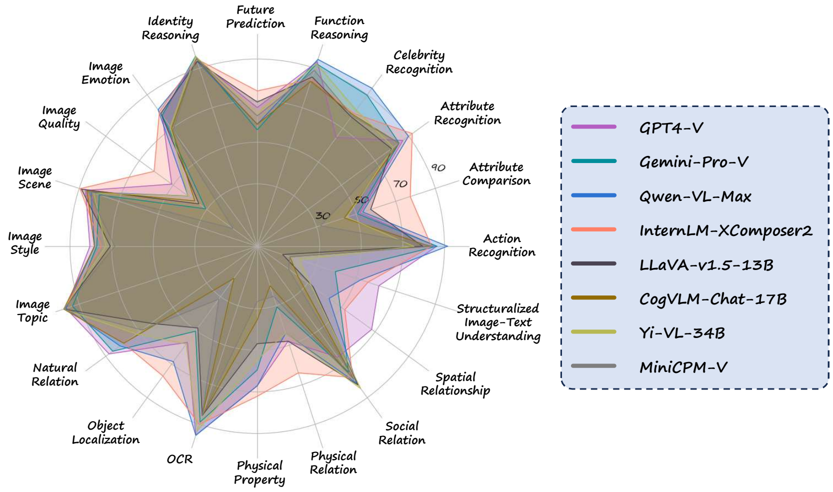


Figure 1: **Results of eight representative large vision-language models (VLMs) across the 20 ability dimensions defined in MMBench-`test`.**

GPT-4v [37], Gemini-Pro-V [44] and LLaVA [33], have demonstrated enhanced capabilities in image content recognition and reasoning within the domain of vision-language models, exhibiting superior performance compared to earlier works. Nevertheless, a large proportion of the early studies [18, 56, 33] tend to emphasize showcasing qualitative examples rather than undertaking comprehensive and quantitative experiments to thoroughly assess their model performance. The lack of quantitative assessment poses a considerable challenge for comparing various models. Recent studies have primarily explored two approaches to conduct quantitative evaluations. The first approach involves utilizing existing public datasets [19, 9] for objective evaluation, while the second approach employs human annotators [49, 48] to perform subjective evaluations. However, it is worth noting that both approaches exhibit some inherent limitations.

A multitude of public datasets, such as VQAv2 [19], COCO Caption [9], GQA [23], and OK-VQA [35], have long served as valuable resources for the quantitative evaluation of VLMs. These datasets offer **objective** metrics, including accuracy, BLEU, CIDEr, *etc*. However, when employed to evaluate more advanced LVLMs, these benchmarks encounter the following challenges. 1. **False Negative Issues**: Most existing evaluation metrics require an exact match between the prediction and the reference target, leading to potential limitations. For instance, in the VQA task, even if the prediction is “bicycle” while the reference answer is “bike”, the existing metric would assign a negative score to the prediction, resulting in a considerable number of false-negative samples. 2. **Lacking Finegrained Analysis**: Current public datasets predominantly focus on evaluating a model’s performance on specific tasks, offering limited insights into the fine-grained capabilities of these models. Thus, they provide insufficient feedback regarding potential directions for future improvements.

Given the aforementioned challenges, recent studies, such as OwlEval [49] and LVLM-eHub [48] propose human-involved **subjective** evaluation strategies, aiming to address existing methods’ limitations by incorporating human judgment and perception in the evaluation process. OwlEval artificially constructs 82 open-ended questions based on images from public datasets and employs human annotators to assess the quality of VLM predictions. Similarly, inspired by FastChat [53], LVLM-eHub develops an online platform where two models are prompted to answer the same question related to an image. A participant then compares the answers provided by two models. Subjective evaluation strategies offer numerous benefits. These include **accurate matching**, where humans can precisely correlate a prediction with the target, even when expressed in different words, and **comprehensive assessment**, where humans are inclined to juxtapose two predictions considering multiple facets. The ultimate score is computed as the mean score across diverse abilities, facilitating a holistic evaluation of the model’s capabilities.

While subjective evaluation allows for a more comprehensive assessment of a VLM, it also introduces new challenges. Firstly, human evaluations are inherently biased. Consequently, it becomes challenging to reproduce the results presented in a work with a different group of annotators. Also, existing subjective evaluation strategies face scalability issues. Employing annotators for model evaluation after each experiment is an expensive endeavor. Moreover, evaluation datasets of small sizes can result in statistical instability. To ensure a robust evaluation, collecting more data becomes necessary, which in turn demands a significant amount of human labor.

In light of the challenges faced by conventional objective and subjective benchmarks, we propose **MMBench**, a systematically designed objective evaluation benchmark to robustly evaluate different abilities of large vision-language models. Currently, MMBench contains over 3000 multiple-choice questions covering 20 different ability dimensions, such as object localization and social reasoning, for evaluating vision-language models. Each ability dimension encompasses over 125 questions, with the quantity of questions per ability maintained at a roughly equal level. The distribution facilitates a balanced and thorough assessment of these abilities. Since some existing VLMs have limited instruction-following capability and cannot directly output choice labels (A, B, C, *etc.*) for MMBench questions, the evaluation based on exact matching may not yield accurate and reasonable conclusions. In order to reduce the number of false-negative samples during answer matching, we employ GPT-4 to match a model's prediction to candidates choices in a multi-choice question and then output the label for the matched choice. We conduct a comparison between GPT-4-based choice matching and human evaluations, and discovered that GPT-4 can accurately match human assessments in 91.5% of cases, demonstrating its good alignment and robustness as a choice extractor. To make the evaluation more robust, we propose a novel evaluation strategy, named **CircularEval** (details in Sec. 4.3). We comprehensively evaluate 21 well-known vision-language models (across different model architectures and scales) on MMBench and report their performance on different ability dimensions. The performance ranking offers a direct comparison between various models and provides valuable feedback for future optimization. In summary, our main contributions are three-fold:

- **Systematically-constructed Dataset**: To thoroughly evaluate the capacity of a VLM, we carefully curated a dataset comprising a total of 3,217 meticulously selected questions, covering a diverse spectrum of 20 fine-grained skills.
- **Robust Evaluation**: We introduce a novel circular evaluation strategy (CircularEval) to improve the robustness of our evaluation process. After that, GPT-4 is employed to match the model's prediction with given choices, which can successfully extract choices even from predictions of a VLM with poor instruction-following capability.
- **Analysis and Observations**: We perform a comprehensive evaluation of a series of well-known vision-language models using MMBench, and the evaluation results can provide insights to the research community for future improvement.

2 Related Work

2.1 Multimodal Datasets

Large-scale VLMs have shown promising potential in multimodal tasks such as complex scene understanding and visual question answering. Though qualitative results so far are encouraging, quantitative evaluation is of great necessity to systematically evaluate and compare the abilities of different VLMs. Recent works have evaluated their models on numerous existing public multi-modality datasets. COCO Caption [9], Nocaps [3], and Flickr30k [51] provide human-generated image captions and the corresponding task is to describe the image content in the form of text. Visual question answering datasets, such as GQA [23], OK-VQA [35], VQAv2 [19], and Vizwiz [20], contain question-answer pairs related to the given image, used to measure the model's ability on visual perception and reasoning. Some datasets provide more challenging question-answering scenarios by incorporating additional tasks. For example, TextVQA [42] proposes questions about text shown in the image, thus involving the OCR task in question-answering. ScienceQA [34] focuses on scientific topics, requiring the model to integrate commonsense into reasoning. Youcook2 [55] replaces images with video clips, introducing additional temporal information. However, the aforementioned datasets are designed on specific domains, and can only evaluate the model's performance on one or several tasks. Besides, different data formats and evaluation metrics across datasets make it more difficult to comprehensively assess a model's capability. Ye et al. [49] constructed OwlEval, an evaluation set encompassing a variety of visual-related tasks, albeit of a limited size. Fu et al. [17] introduced MME, which assesses a VLM's capabilities from various perspectives at a small scale. Diverging from prior works, in this paper, we present a novel multimodal benchmark, MMBench. We also devise a suite of evaluation standards aimed at ensuring the stability and accuracy of the evaluation results.

2.2 Multimodal Models

Building upon the success of Large Language Models (LLMs) such as GPTs [41, 7, 40], LLaMA [46], and Vicuna [53], recent advancements have been made in multimodal models. Flamingo [4], an early attempt at integrating LLMs into vision-language pretraining, has made significant strides. To condition effectively on visual features, it incorporates several gated cross-attention dense blocks within pretrained language encoder layers. OpenFlamingo [4] offers an open-source version of this model. BLIP-2 [28] introduces a Querying Transformer (Q-former) to bridge the modality gap between the frozen image encoder and the large language encoder. Subsequently, InstructBLIP [11] extends BLIP-2 [28] with vision-language instruction tuning, achieving superior performance. MiniGPT-4 [56] attributes the prowess of GPT-4 [37] to advanced LLMs and proposes the use of a single projection layer to align the visual representation with the language model. LLaVA [33] also utilizes GPT-4 to generate instruction-following data for vision-language tuning. The learning paradigm and the multimodal instruction tuning corpus proposed by LLaVA are widely adopted by subsequent works [32, 8, 2, 10]. During the instruction tuning, Low-Rank Adaptation (LoRA [22]) has been adopted by recent works [49, 12, 10] on language models to achieve better performance on multimodal understanding. In the realm of proprietary models, the APIs of multiple powerful VLMs have also been made publicly available to prosper downstream applications, including GPT-4v [37], Gemini-Pro-V [44], and Qwen-VL-Max [6]. After conducting a thorough evaluation of these models on the proposed MMBench, we offer insights for future multimodal research.

3 The construction of MMBench

Three characteristics differentiate MMBench from existing benchmarks for multi-modality understanding: i) MMBench adopts images / problems from various sources to evaluate diversified abilities in a hierarchical taxonomy; ii) MMBench performs rigorous quality control to ensure the correctness and validity of testing samples; iii) MMBench is a bilingual multi-modal benchmark and enables an apple-to-apple comparison of VLM performance under English and Chinese contexts. Below we will delve into more details of the construction of MMBench.

3.1 The Hierachical Ability Taxonomy of MMBench

Human possess remarkable perception and reasoning capabilities. These abilities have been crucial in human evolution and serve as a foundation for complex cognitive processes. Perception refers to gathering information from sensory inputs, while reasoning involves drawing conclusions based on this information. Together, they form the basis of most tasks in the real world, including recognizing objects, solving problems, and making decisions [36, 16]. In pursuit of genuine general artificial intelligence (AGI), vision-language models (VLMs) are also expected to exhibit strong perception and reasoning abilities. Therefore, we adopt **Perception** and **Reasoning** as level-1 (**L-1**) abilities in our taxonomy. After that, we incorporate more fine-grained ability dimensions into the taxonomy, and categorize them into six **L-2** and twenty **L-3** ability dimensions. We display the ability taxonomy in Figure 2 and you can find detailed definitions of each fine-grained ability in the Appendix.

3.2 Data Collection and Quality Control

Question Collection. In MMBench, we collect vision-language QAs in the format of multiple-choice problems for each L-3 ability. A problem P_i corresponds to a quadruple (Q_i, C_i, I_i, A_i). Q_i denotes the question, C_i represents a set with n ($2 \leq n \leq 4$) choices $c_1, c_2, ..., c_n$, I_i corresponds to the image associated with the question, and A_i is the correct answer. The data — including images, choices, and questions — are manually collected from multiple sources by a group of volunteers. For each **L-3** ability, we first set an example by compiling $10 \sim 20$ multiple-choice questions. Then we enlist the volunteers, all of whom are undergraduate or graduate students from various disciplines, to expand the problem set. The expansion is based on the ability definition and potential data sources, which include both public datasets and the Internet. According to the statistics, more than 80% of questions in MMBench are collected from the Internet. For the remaining 20% samples, the

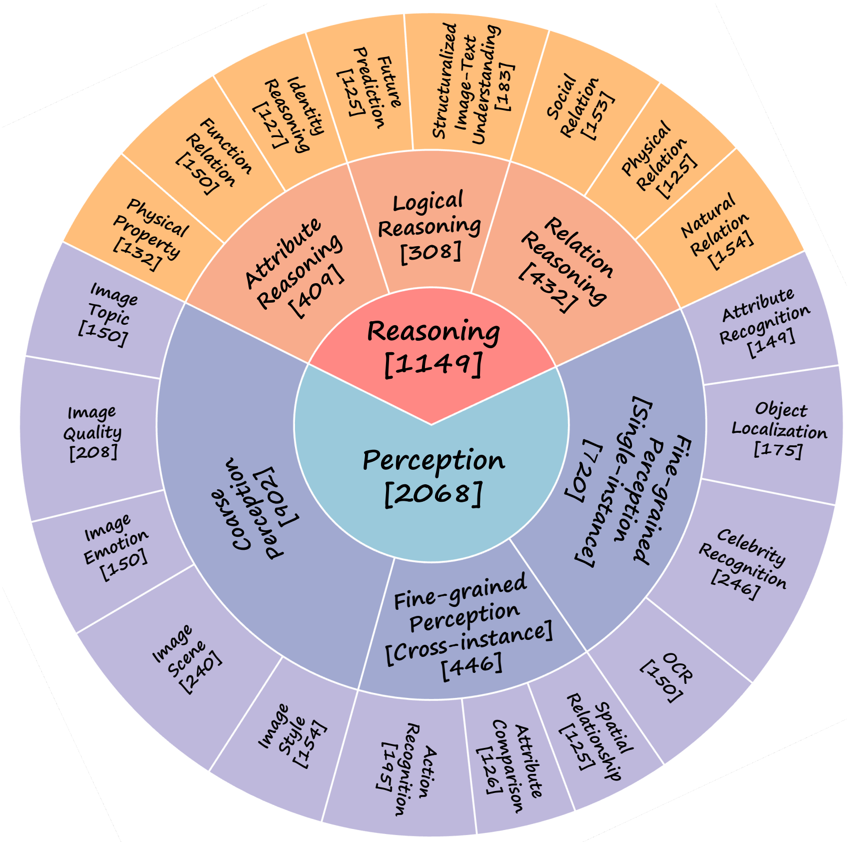
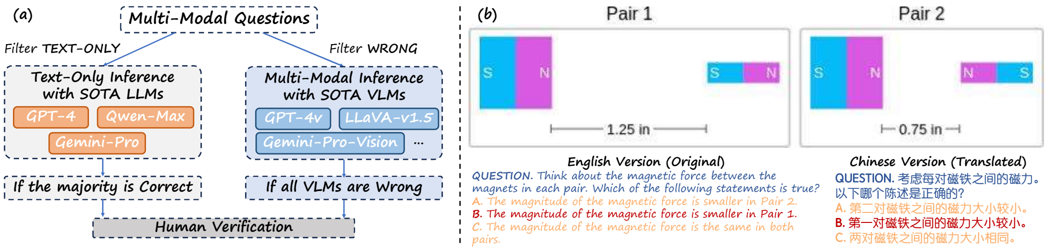


Figure 2: **Ability dimensions in MMBench.** Currently, MMBench incorporates three levels of ability dimensions, encompassing 20 distinct leaf abilities.



QUESTION. Think about the magnetic force between the magnets in each pair. Which of the following statements is true?
A. The magnitude of the magnetic force is smaller in Pair 2.
B. The magnitude of the magnetic force is smaller in Pair 1.
C. The magnitude of the magnetic force is the same in both pairs.

QUESTION. 考虑每对磁铁之间的磁力。以下哪个陈述是正确的?
A. 第二对磁铁之间的磁力大小较小。
B. 第一对磁铁之间的磁力大小较小。
C. 两对磁铁之间的磁力大小相同。

Figure 3: **The construction of MMBench.** (a). The quality control strategies adopted in MMBench; (b) An illustration of questions in MMBench-CN.

images are gathered from the validation set of public datasets (if they exist) while the questions are self-constructed, which is not supposed to be used for training. In the Appendix, we list data sources used in collection and provide visualization of samples corresponding to each **L-3** ability.

Quality Control. Raw data collected from volunteers may include wrong or unqualified samples. During investigation, we find that there exist two major patterns for such samples: i) the answer to the question can be inferred with **text-only** inputs, which makes it inappropriate for evaluating the multimodal understanding capability of VLMs; ii) the sample is simply **wrong**, either with a flawed question, choices, or an incorrect answer. We design two strategies to filter those low-quality samples, which is visualized in Figure 3(a). We adopt 'majority voting' to detect **text-only** samples: data samples are inferred with state-of-the-art LLMs (GPT-4 [37], Gemini-Pro [44], *etc.*). If more than half of the LLMs can answer the question correctly with text-only inputs, the question will be manually verified and then removed if it is unqualified. To detect **wrong** samples, we also implement an automatic filtering mechanism. We select several state-of-the-art VLMs (including both open-

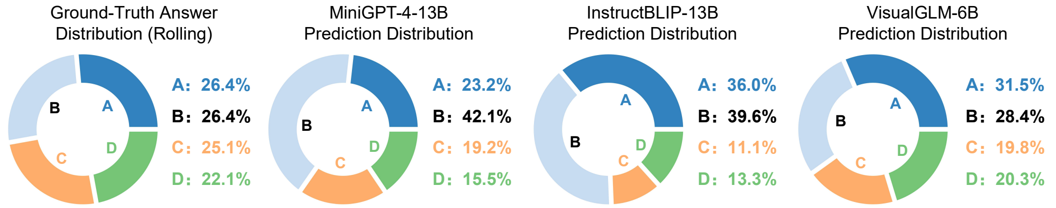


Figure 4: **The choice distribution of ground-truth answers and predictions of sample VLMs (all *CircularEval* records).** Since there exist questions with only 2/3 choices in MMBench, the choice distribution of ground-truth is not exactly even.

source and proprietary ones), to answer all questions in MMBench . If all VLMs fail to answer the question correctly, we consider this question potentially problematic. Such questions will be manually checked and excluded if they are actually wrong. The quality control paradigm helps us to construct high-quality datasets and can also be used to clean other existing benchmarks.

MMBench-CN. We further convert the curated MMBench into a Chinese version. During the process, all content in questions and choices are translated to Chinese based on GPT-4, except for proper nouns, symbols, and code. All those translations are verified by humans to ensure the validity. MMBench-CN enables an apple-to-apple comparison of VLM performance under English and Chinese contexts. An example in MMBench-CN is illustrated in Figure 3(b).

3.3 MMBench Statistics

Data Statistics. In the present study, we have gathered a total of 3,217 data samples spanning across 20 distinct **L-3** abilities. We depict the problem counts of all the 3 levels of abilities in Figure 2. To ensure a balanced and comprehensive evaluation for each ability, we try to maintain an even distribution among problems associated with different abilities during data collection, with at least 125 samples for each **L-3** category.

Data Splits. We follow the standard practice in previous works [35] to split MMBench into `dev` and `test` subsets at a ratio of 4:6. For the `dev` subset, we make all data samples publicly available along with the ground truth answers for all questions. For the `test` subset, only the data samples are released, while the ground truth answers remain confidential. To obtain the test subset evaluation results, one needs to submit the predictions to MMBench evaluation server.

4 Evaluation Strategy

In MMBench, we propose a new strategy that yields robust evaluation results with affordable costs. To deal with the free-form outputs of VLMs, we propose utilizing state-of-the-art LLMs as a helper for choice extraction. We conduct extensive experiments to study the LLM-involved evaluation procedure. The results well support the effectiveness of GPT-4 as a choice extractor. We further adopt a new evaluation strategy named **CircularEval**, which feeds a question to a VLM multiple times (with shuffled choices) and checks if a VLM succeeds in all attempts. With **CircularEval**, we deliver a rigorous evaluation and more effectively display the performance gap between VLMs.

4.1 LLM-involved Choice Extraction

In our initial attempts to evaluate on MMBench questions, we observed that the instruction-following capabilities of VLMs can vary significantly. Though problems are presented as clear multiple-choice questions with well-formatted options, many VLMs still output the answers in free-form text[2], especially for VLMs that have not been trained with multiple-choice questions or proprietary VLMs for general purposes (GPT-4v, Qwen-VL-Max, *etc.*). Extracting choices from free-form predictions is straight-forward for human beings, but might be difficult with rule-based matching. To this end, we design a universal evaluation strategy for all VLMs with different instruction-following capabilities:

---

[2] For example, the model output can be the meaning of choice "A" rather than "A".

Table 1: **Statistics of IF capabilities of VLMs.** We report the heuristic matching success rate of VLMs, and the accuracy before and after LLM-based choice extraction. In 'X+Y', X denotes the matching-based accuracy, Y indicates the gain of using LLM as the choice extractor.

Model Name	Match Rate	DEV Acc	Model Name	Match Rate	DEV Acc
MiniGPT4-7B	85.7	47.9 +8.8	MiniGPT4-13B	84.8	52.1 +8.7
InstructBLIP-7B	93.6	57.1 +4.3	InstuctBLIP-13B	93.7	58.4 +5.6
IDEFICS-9B-Instruct	96.6	58.4 +1.5	Qwen-VL-Chat	93.8	73.3 +3.6
MiniCPM-V	95.2	70.9 +4.5	VisualGLM-6B	64.8	39.9 +23.2
GPT-4v	91.8	81.5 +3.6	GeminiProVision	97.5	81.8 +0.8
Qwen-VL-Plus	77.4	64.5 +15.0	Qwen-VL-Max	96.0	82.0 +3.2

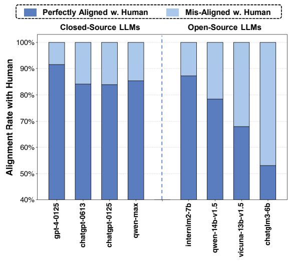
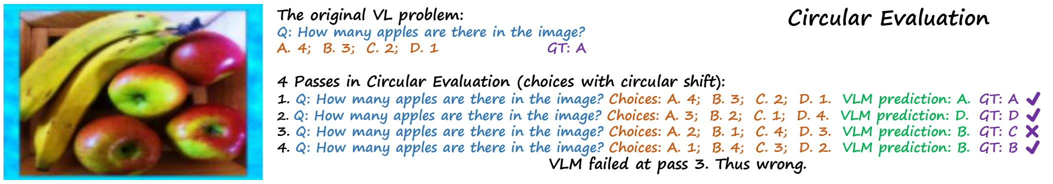


Figure 5: **Alignment rates between human and different LLMs.** 'chatgpt' is 'gpt-3.5-turbo'. Open-source LLMs are 'chat' variants.



The original VL problem:
Q: How many apples are there in the image?
A. 4; B. 3; C. 2; D. 1 GT: A

4 Passes in Circular Evaluation (choices with circular shift):
1. Q: How many apples are there in the image? Choices: A. 4; B. 3; C. 2; D. 1. VLM prediction: A. GT: A ✓
2. Q: How many apples are there in the image? Choices: A. 3; B. 2; C. 1; D. 4. VLM prediction: D. GT: D ✓
3. Q: How many apples are there in the image? Choices: A. 2; B. 1; C. 4; D. 3. VLM prediction: B. GT: C ✗
4. Q: How many apples are there in the image? Choices: A. 1; B. 4; C. 3; D. 2. VLM prediction: B. GT: B ✓

VLM failed at pass 3. Thus wrong.

Figure 6: **CircularEval strategy.** In **CircularEval**, a problem is tested multiple times with circular shifted choices and the VLM needs to succeed in all testing passes. In this example, the VLM failed in pass 3 and thus considered failed the problem.

Step 1. Matching Prediction. Initially, we attempt to extract choices from VLM predictions using heuristic matching. We aim to extract the choice label (e.g., A, B, C, D) from the VLM's output. If successful, we use this as the prediction. If not, we attempt to extract the choice label using an LLM.

Step 2. Matching LLM's output. If step 1 fails, we try to extract the choice with LLMs (**gpt-4-0125** by default). We first provide ChatGPT with the question, choices, and model prediction. Then, we request it to align the prediction with one of the given choices, and subsequently produce the label of the corresponding option. If the LLM finds that the model prediction is significantly different from all choices, we ask it to return a pseudo choice 'Z'. In experiments, we find that for almost all cases we encountered, the LLM can output a valid choice according to the instruction. For each sample, we compare the model's label prediction (after GPT's similarity readout) with the actual ground truth label. If the prediction matches the label, the test sample is considered correct.

4.2 LLM as the Choice Extractor: A Feasibility Analysis

Instruction following (IF) capabilities of VLMs vary a lot. We conduct pilot experiments to study the effectiveness of LLMs as the choice extractor. As a first step, we perform single-pass inference on all MMBench questions with VLMs in our evaluation core set (defined in Sec. 5.2). While there exist VLMs that perfectly follow the multiple-choice format and achieve high success rates ($> 99\%$) in heuristic matching, all proprietary models and a significant proportion of open-source VLMs failed to generate well-formatted outputs. In Table 1, we list the success rates of different VLMs in heuristic matching[3]. Among all VLMs, VisualGLM achieves the lowest matching success rate, which is merely 65%. For those VLMs, incorporating LLMs as the choice extractor leads to significant change in the final accuracy. Another noteworthy thing is that the IF capability and the overall multimodal understanding capability is not necessarily correlated. For example, OpenFlamingo v2 [4] demonstrates top IF capability among all VLMs, while also achieving one of the worst performances on MMBench (Table 3).

---

[3] VLMs that achieve $> 99\%$ matching rates are not listed, including LLaVA series, Yi-VL series, mPLUG-Owl2, OpenFlamingo v2, and CogVLM-Chat.

Quality and stability of LLM Choice Extractors. For VLM predictions that cannot be parsed by heuristic matching, we adopt GPT-4 as the choice extractor. To validate its efficacy, we first build a subset of the inference records. Each item in the set is a pair of questions and VLM predictions, which cannot be parsed by step-1 matching. We sample 10% of those hard examples ($\sim$ 420 samples), and ask volunteers to perform manual choice extraction on these data samples. Such annotations enable us to validate the choice extraction of LLMs, by measuring their alignment rates with humans.

Figure 5 reports the alignment rates (extracted choices are exactly the same) between LLMs and humans. We find that a great number of LLMs can complete the task well and achieve decent alignment rate with human. Among proprietary LLMs, GPT-4 achieves the highest level of alignment rate, which is 91.5%, while GPT-3.5-Turbo and Qwen-Max achieve around 85%. Open-source LLMs achieve more diversified performance on the choice matching task. InternLM2-7B [45] achieves an 87% alignment rate and significantly outperforms other open-source LLMs and GPT-3.5-Turbo. In the following experiments, we adopt **gpt-4-0125** as the choice extractor due to its superior alignment capability. Meanwhile, we also note that the slight difference in top-performing LLMs' alignment rates has little effect on the quantitative performance of VLMs.

4.3 CircularEval Strategy

In MMBench, the problems are presented as multiple-choice questions. Such formulation poses an evaluation challenge: random guessing will lead to $\sim$25% Top-1 accuracy for 4-choice questions, potentially reducing the discernible performance differences among VLMs. Besides, we noticed that VLMs may prefer to predict a certain choice among all given choices (Figure 4), which further amplifies the bias in evaluation. To this end, we introduce a more robust evaluation strategy termed **Circular Evaluation** (or **CircularEval**). Under this setting, each question is fed to a VLM N times (N is the number of choices). Each time, circular shifting is applied to the choices and the answer to generate a new prompt for VLMs (example in Figure 6). A VLM is considered successful in solving a question only if it correctly predicts the answer in all circular passes. In practice, once a VLM fails on a circular passes, there is no need to infer the remaining passes, which makes the actual cost of CircularEval less than $N\times$ under practical scenarios. CircularEval can achieve a good trade-off between robustness and cost.

5 Evaluation Results

5.1 Experimental Setup

For the main results, we evaluate various models belonging to three major categories on MMBench: (a) *Text-Only* GPT-4 [37]; (b) *Open-Source VLMs* including model variants of OpenFlamingo [4], MiniGPT4 [56], InstructBLIP [11], LLaVA [32], IDEFICS [26], CogVLM [47], Qwen-VL [6], Yi-VL [2], mPLUG-Owl [50], InternLM-XComposer [12], and MiniCPM-V [39]; (c) *Proprietary VLMs* including Qwen-VL-[Plus/Max] [6], Gemini-Pro-V [44], and GPT-4v [37]. For a fair comparison, we adopt the zero-shot setting to infer MMBench questions with all VLMs, based on the same prompt. For all VLMs, open-ended generation is adopted to obtain the prediction, and '**gpt-4-0125**' is used as the choice extractor. In the Appendix, we provide detailed information regarding the architecture and the parameter size for all Open-Source VLMs evaluated in this paper, as well as additional results for more VLMs under various settings. We conduct all the evaluation with VLMEvalKit [14].

5.2 Main Results

CircularEval *vs.* VanillaEval. Before delving deeper into concrete evaluation results, we first compare our **CircularEval** (infer a question over multiple passes, consistency as a must) with **VanillaEval** (infer a question only once). In Table 2, we present the results with two evaluation strategies on MMBench-`dev`. For most VLMs, switching from VanillaEval to CircularEval leads to a significant drop in model accuracy. In general, comparisons under CircularEval can reveal a more significant performance gap between different VLMs. LLaVA-v1.5-13B outperforms its 7B counterpart by 2.1% Top-1 accuracy under VanillaEval, while a much larger performance gap (4.7% Top-1) is observed under CircularEval. As a special case, the performance of OpenFlamingo v2 drops from 36.7% to only 2.6% when we move from VanillaEval to CircularEval. CircularEval is such a challenging setting that it even makes state-of-the-art proprietary VLMs (GPT-4v, Qwen-VL-Max,

Table 2: **CircularEval** *vs.* **VanillaEval.** We report the **CircularEval** Top-1 accuracy and accuracy drop (compared to **VanillaEval**) of all VLMs on MMBench-`dev`.

VLM	Circular	Acc Change	VLM	Circular	Acc Change	VLM	Circular	Acc Change
MiniGPT4-7B	32.7%	**-24.1%**	**MiniGPT4-13B**	37.5%	**-23.2%**	**Yi-VL-6B**	65.6%	**-9.8%**
InstructBLIP-7B	37.4%	**-24.0%**	**InstructBLIP-13B**	40.9%	**-23.0%**	**Yi-VL-34B**	68.2%	**-9.5%**
LLaVA-v1.5-7B	62.5%	**-11.2%**	**LLaVA-v1.5-13B**	67.2%	**-8.6%**	**MiniCPM-V**	64.8%	**-10.6%**
IDEFICS-9B-Instruct	37.2%	**-22.6%**	**LLaVA-InternLM2-20B**	72.8%	**-7.0%**	**Qwen-VL-Plus**	62.9%	**-16.6%**
VisualGLM-6B	36.1%	**-27.0%**	**CogVLM-Chat-17B**	62.4%	**-15.6%**	**Qwen-VL-Max**	76.4%	**-8.7%**
Qwen-VL-Chat	59.5%	**-17.4%**	**mPLUG-Owl2**	63.5%	**-8.7%**	**Gemini-Pro-V**	70.9%	**-11.7%**
OpenFlamingo v2	2.6%	**-34.1%**	**InternLM-XComposer2**	79.1%	**-4.7%**	**GPT-4v**	74.3%	**-10.8%**

etc.) suffer from $\sim$10% Top-1 accuracy drops. In the following experiments, we adopt the more rigorous and well-defined **CircularEval** as our default evaluation paradigm.

We exhaustively evaluate all VLMs on all existing leaf abilities of MMBench. In Table 3, we report the models' overall performance and the performance in six **L-2** abilities on the `test` split, namely Coarse Perception (**CP**), Fine-grained Perception (single-instance, **FP-S**; cross-instance, **FP-C**), Attribute Reasoning (**AR**), Logic Reasoning (**LR**), and Relation Reasoning (**RR**).[4] The results offer valuable insights into the individual strengths and limitations of each VLM in multi-modal understanding.

Performance on MMBench-`test`. We first conduct a sanity check by inferring MMBench questions with GPT-4, using text-only inputs. After conducting the rigorous quality control paradigm in Sec. 3.2, GPT-4 demonstrates a random-level overall accuracy. Among open-source VLMs, InternLM-XComposer2 [12] achieves the best performance and surpass other open-source or proprietary models by a large margin, *w.r.t.* the overall score, demonstrating its superior ability in multimodal understanding. After that, models adopting the architecture of LLaVA [33] (LLaVA series and Yi-VL series) also showcase strong overall performance, which is just inferior to the state-of-the-art closed-source GPT-4v and Qwen-VL-Max. With a small parameter size ($\leq$ 3B), MiniCPM-V achieves over 60% Top-1 accuracy, highlighting the potential of small-scale VLMs. Models including MiniGPT, IDEFICS, VisualGLM, and InstructBLIP demonstrate significantly inferior performance compared to other VLMs, while OpenFlamingo v2 shows random-level performance due to the lack of instruction tuning.

LLM plays a vital role. From the evaluation results, we find that the large language model (LLM) adopted plays a vital role in the VLM performance. For instance, all LLaVA series VLMs (v1.5-7B, v1.5-13B, InternLM2-20B) adopt the same vision backbone and are trained with the same multimodal corpus, while switching the LLM from Vicuna-v1.5 [53] to the more powerful InternLM2-20B [45] leads to steady improvement across all L-2 capabilities (especially significant for reasoning tasks). The scaling also holds for variants with different sizes from the same LLM family. By adopting the 13B variant of Vicuna rather than the 7B variant, VLMs in the MiniGPT, InstructBLIP, and LLaVA v1.5 series outperform their 7B counterparts by 8.3%, 1.5%, and 3.5% overall Top-1 accuracies on the MMBench-`test` split, respectively.

Performance on MMBench-CN. In Figure 7, we compare the performance of different VLMs on MMBench and MMBench-CN. Most VLMs display a lower performance on MMBench-CN compared to the results on MMBench, except OpenFlamingo v2, VisualGLM, and Qwen-VL-Plus. The difference may be attributed to the unbalanced English and Chinese corpora used in the pretraining and instruction-tuning of VLMs and their corresponding LLMs. We notice that most top-performing VLMs on MMBench also display outstanding performance under the bilingual context. The largest EN-CN performance gap for models that achieve 70+% Top-1 accuracy on MMBench is a mere 2%, For InternLM-XComposer2, the accuracy only drops by less than 1% when evaluated on MMBench-CN. Such an advantage can be attributed to utilizing LLMs with better bilingual capabilities or tuning the VLM with more balanced cross-language multimodal corpora.

5.3 Fine-grained Analysis

In this section, we present more fine-grained analysis based on the evaluation results.

---

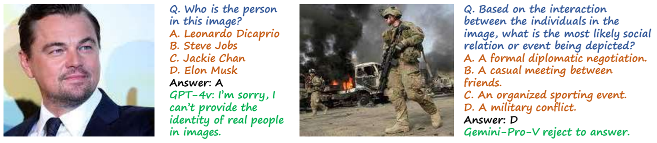
[4] Please refer to the appendix for more fine-grained results and MMBench-`dev` split results.

Table 3: **CircularEval results on MMBench `test` set (L-2 abilities).** Abbreviations adopted: LR for Logical Reasoning; AR for Attribute Reasoning; RR for Relation Reasoning; FP-C for Fine-grained Perception (Cross Instance); FP-S for Fine-grained Perception (Single Instance); CP for Coarse Perception. Models are sorted by the ascending order of overall accuracy (intra-group). Open-source models tagged with * incorporate in-house data in model training.

Model	**Overall**	**CP**	**FP-S**	**FP-C**	**AR**	**LR**	**RR**
Large Language Models							
GPT-4-Turbo (0125) [37]	2.9%	0.6%	1.2%	4.1%	3.7%	4.9%	7.4%
OpenSource VLMs							
OpenFlamingo v2 [4]	2.3%	1.1%	3.5%	1.5%	5.3%	0.0%	2.7%
MiniGPT4-7B [56]	30.5%	37.0%	31.8%	17.2%	49.8%	9.2%	25.6%
IDEFICS-9B-Instruct [26]	35.2%	48.3%	31.3%	29.6%	47.8%	11.4%	25.2%
VisualGLM-6B [13]	35.4%	40.2%	38.5%	26.2%	47.8%	19.6%	29.5%
InstructBLIP-7B [11]	38.3%	46.7%	39.0%	31.8%	55.5%	8.7%	31.0%
MiniGPT4-13B [56]	38.8%	44.6%	42.9%	23.2%	64.9%	8.2%	32.9%
InstructBLIP-13B [11]	39.8%	47.2%	42.9%	21.0%	60.4%	12.5%	38.8%
Qwen-VL-Chat* [6]	60.9%	68.5%	67.7%	50.2%	78.0%	37.0%	45.7%
MiniCPM-V [39]	61.4%	65.6%	69.4%	51.3%	70.6%	35.3%	59.7%
LLaVA-v1.5-7B [32]	63.4%	70.0%	68.0%	57.7%	77.6%	33.2%	56.2%
mPLUG-Owl2 [50]	63.5%	68.1%	69.1%	55.8%	78.4%	37.0%	57.0%
CogVLM-Chat-17B [47]	63.6%	72.8%	66.6%	55.4%	71.4%	33.7%	62.0%
Yi-VL-6B* [2]	65.5%	72.8%	72.9%	56.2%	75.5%	41.3%	55.4%
LLaVA-v1.5-13B [32]	66.9%	73.1%	72.4%	60.3%	75.5%	35.9%	65.5%
Yi-VL-34B* [2]	68.4%	72.0%	78.0%	54.7%	81.2%	38.6%	68.2%
LLaVA-InternLM2-20B [10]	72.3%	78.3%	76.6%	68.2%	78.4%	46.2%	69.4%
InternLM-XComposer2* [12]	78.1%	80.4%	83.5%	73.0%	83.7%	63.6%	74.4%
Proprietary VLMs							
Qwen-VL-Plus [6]	64.6%	66.5%	79.1%	50.2%	73.9%	42.9%	57.8%
Gemini-Pro-V [44]	70.2%	70.0%	78.9%	65.9%	82.9%	46.2%	65.9%
GPT-4v [37]	74.3%	77.6%	73.8%	71.5%	85.3%	63.6%	68.6%
Qwen-VL-Max [6]	75.4%	74.8%	87.2%	67.0%	85.3%	54.9%	70.5%

Table 4: **'Upper-bound' Acc Estimation for Proprietary VLMs.**

Model	MMBench-`test`	Upper Bound
GPT-4v	74.3	76.2
Gemini-Pro-V	70.2	72.6
Qwen-VL-Max	75.4	75.5



Q. Who is the person in this image?
A. Leonardo Dicaprio
B. Steve Jobs
C. Jackie Chan
D. Elon Musk
Answer: A
GPT-4v: I'm sorry, I can't provide the identity of real people in images.

Q. Based on the interaction between the individuals in the image, what is the most likely social relation or event being depicted?
A. A formal diplomatic negotiation.
B. A casual meeting between friends.
C. An organized sporting event.
D. A military conflict.
Answer: D
Gemini-Pro-V reject to answer.

Figure 8: **Content Moderation Cases of Proprietary VLMs.**

Content Moderation of Proprietary VLMs. When we take an in-depth look at the predictions of proprietary VLMs, we notice that all of them apply explicit content moderation. GPT-4v, Gemini-Pro-V, and Qwen-VL-Max reject answering in 1.8%, 1.6%, and 0.1% of cases across all CircularEval passes in MMBench, respectively. 74% of questions rejected by GPT-4v are related to celebrity recognition (Figure 8), while no obvious rejection pattern is observed for Gemini-Pro-V. Under CircularEval, such moderation has a negative impact on the evaluated accuracy. To estimate an **upper-bound** performance, we assume that VLMs can perfectly answer all rejected questions and re-calculate the accuracy. Table 4 shows that the content moderation policy affects the MMBench-`test` accuracy by up to 2.4%, which is not a significant change.

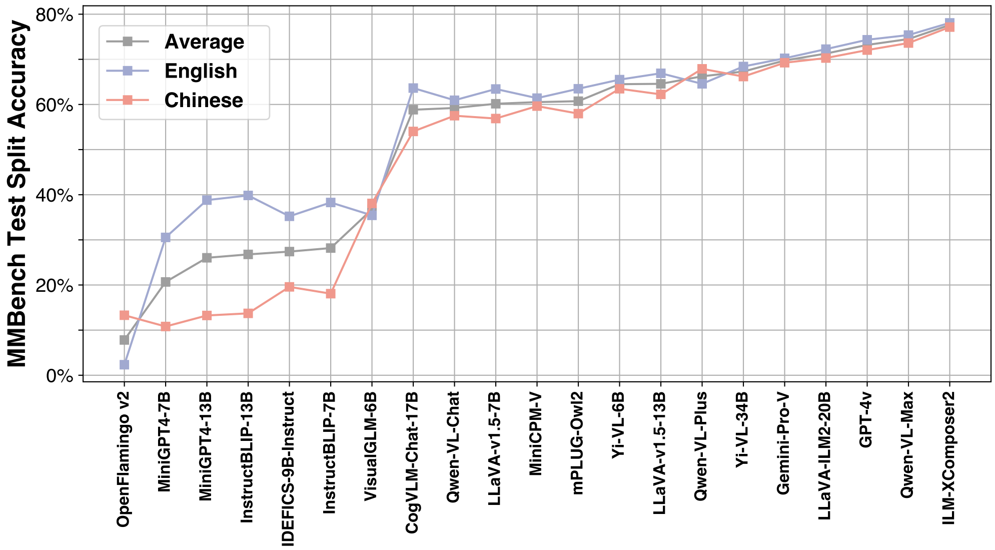


Figure 7: **The performance on the `test` split of MMBench and MMBench-CN.** Models are sorted with the ascending order of average performance. ILM stands for InternLM.

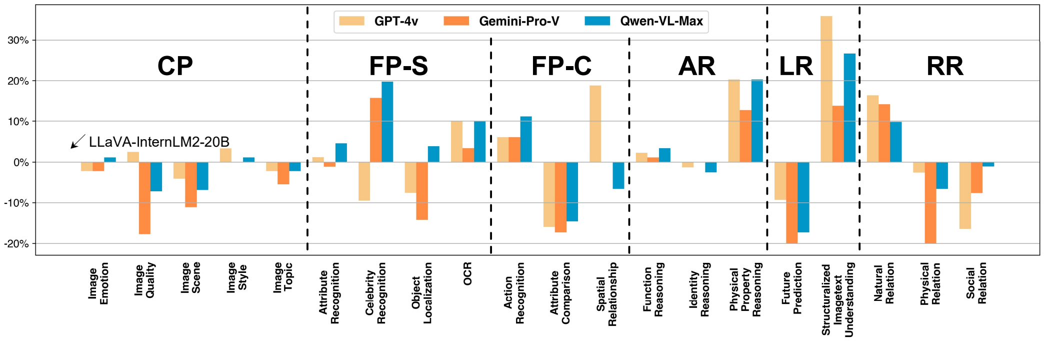
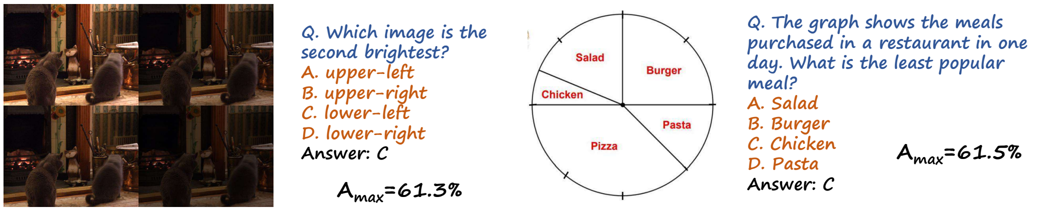


Figure 9: **Proprietary VLMs *vs.* Open-Source ones at a fine-grained level.**

Proprietary *vs.* Open-Source: What is the gap? Compared to the varied performance of open-source VLMs, most proprietary models demonstrate competitive performance on MMBench. This raises a question we care about: are proprietary models generally more powerful, or do each kind of model display unique strengths and weaknesses across different types of ability? To answer this question, we perform a fine-grained comparison of three proprietary VLMs and LLaVA-InternLM2-20B, the top-performing model trained on open-source datasets only, and visualize the result in Figure 9. We observe that proprietary models significantly outperform the open-source ones under two major scenarios: i) **Structuralized image-text understanding**, which requires VLMs to understand complex codes, tables, diagrams, or layouts. ii) **Tasks requiring external knowledge to solve**, which correspond to abilities including celebrity recognition, physical property reasoning, natural relation reasoning, *etc.* Meanwhile, proprietary VLMs do not display advantages on tasks corresponding to other perception or reasoning capabilities.

Hard cases in MMBench. For most VLMs, the fine-grained accuracies vary a lot across different ability categories. To provide insights for future VLM optimization, we find the maximum accuracy (A_{max}) across all evaluated VLMs on each L-3 capability. Samples belonging to L-3 capabilities with the lowest A_{max} are visualized in Figure 10. Generally, we find that all existing VLMs have the following limitations: 1. Poor at recognizing the low-level features on visual inputs, *i.e.*, they cannot accurately recognize and compare the brightness, sharpness, contrast ratio, or artifacts of images. 2. Difficulty in understanding structuralized visual inputs like tables, diagrams, or layouts, even for relatively simple cases like Figure 10(b); 3. Perform badly on recognizing or reasoning about the inter-object spatial relationships, either in 2D or 3D space.



Q. Which image is the second brightest?
A. upper-left
B. upper-right
C. lower-left
D. lower-right
Answer: C

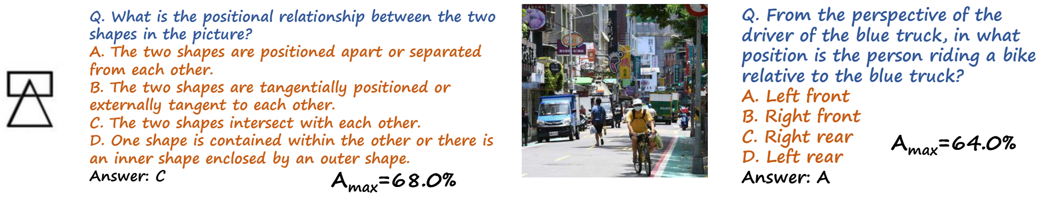
A_{max}=61.3%

Q. The graph shows the meals purchased in a restaurant in one day. What is the least popular meal?
A. Salad
B. Burger
C. Chicken
D. Pasta
Answer: C

A_{max}=61.5%

(a). Image Quality

(b). Structralized Image-Text Understanding



Q. What is the positional relationship between the two shapes in the picture?
A. The two shapes are positioned apart or separated from each other.
B. The two shapes are tangentially positioned or externally tangent to each other.
C. The two shapes intersect with each other.
D. One shape is contained within the other or there is an inner shape enclosed by an outer shape.
Answer: C

A_{max}=68.0%

Q. From the perspective of the driver of the blue truck, in what position is the person riding a bike relative to the blue truck?
A. Left front
B. Right front
C. Right rear
D. Left rear
Answer: A

A_{max}=64.0%

(c). Spatial Relationship

(d). Physical Relation Reasoning

Figure 10: **Hard examples that belong to the 4 L-3 abilities with lowest A_{max}.** All VLMs have made the wrong prediction for the visualized examples under CircularEval.

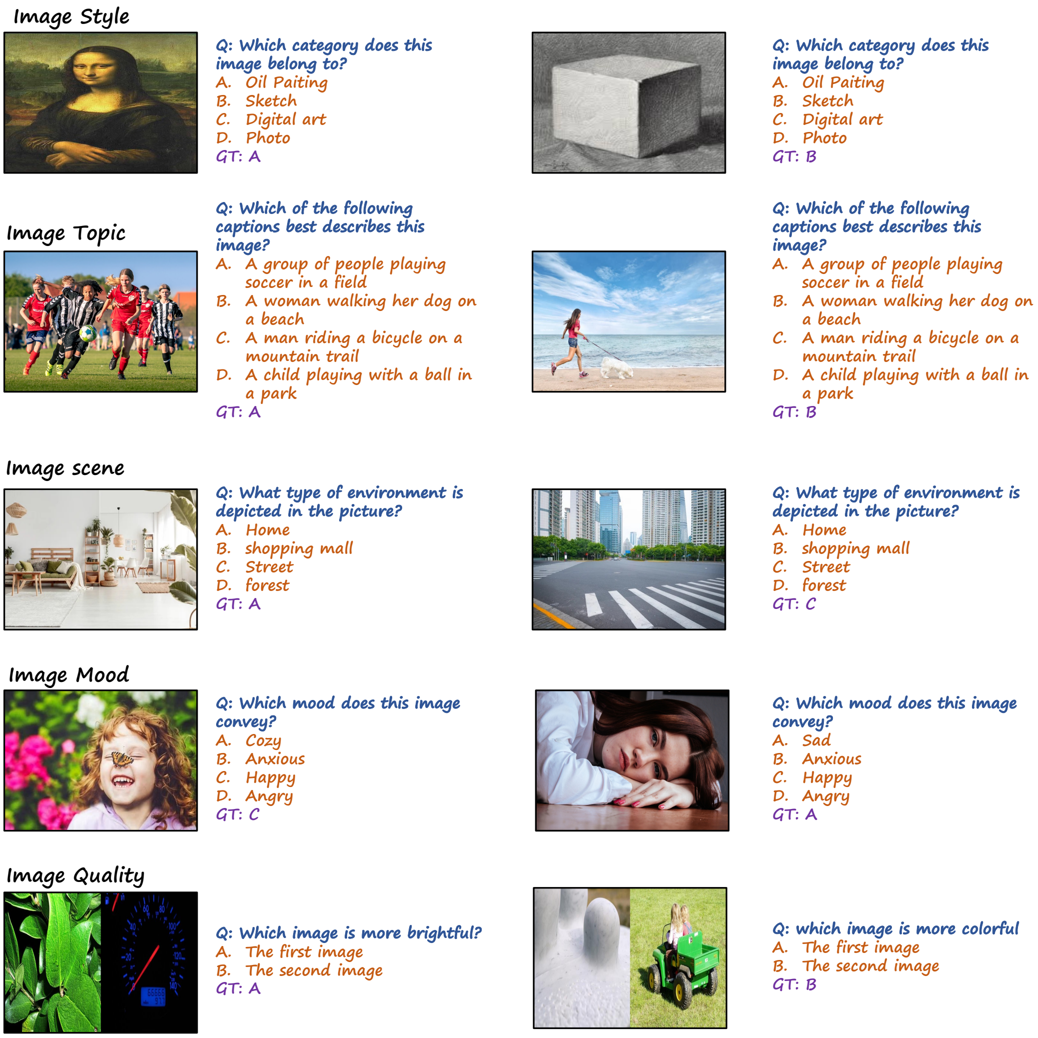
6 Conclusion

We introduce MMBench, a multi-modality benchmark that performs objective evaluation for VLMs with over 3,000 multiple-choice questions covering 20 ability dimensions. To produce robust and reliable evaluation results, we introduce a new evaluation strategy named **CircularEval**. The strategy is much stricter than the vanilla 1-pass evaluation and can yield reliable evaluation results at an affordable cost. Considering the limited instruction following ability of some VLMs, to yield more accurate evaluation results, we additionally adopt LLMs to extract choices from the model's predictions. We comprehensively evaluate over 20 mainstream VLMs on MMBench, covering different architectures and parameter sizes. The evaluation results provide valuable insights for future improvements.

A More Details about the Data

In this section, we begin by providing a detailed definition of each leaf ability (L-3) and present a collection of visualization samples that are directly related to each leaf ability. Then, we enumerate all the data sources that were utilized in the construction of MMBench.

A.1 Definition about Each Leaf Ability



Q: Which category does this image belong to?
A. Oil Paiting
B. Sketch
C. Digital art
D. Photo
GT: A

Q: Which category does this image belong to?
A. Oil Paiting
B. Sketch
C. Digital art
D. Photo
GT: B

Q: Which of the following captions best describes this image?
A. A group of people playing soccer in a field
B. A woman walking her dog on a beach
C. A man riding a bicycle on a mountain trail
D. A child playing with a ball in a park
GT: A

Q: Which of the following captions best describes this image?
A. A group of people playing soccer in a field
B. A woman walking her dog on a beach
C. A man riding a bicycle on a mountain trail
D. A child playing with a ball in a park
GT: B

Q: What type of environment is depicted in the picture?
A. Home
B. shopping mall
C. Street
D. forest
GT: A

Q: What type of environment is depicted in the picture?
A. Home
B. shopping mall
C. Street
D. forest
GT: C

Q: Which mood does this image convey?
A. Cozy
B. Anxious
C. Happy
D. Angry
GT: C

Q: Which mood does this image convey?
A. Sad
B. Anxious
C. Happy
D. Angry
GT: A

Q: Which image is more brightful?
A. The first image
B. The second image
GT: A

Q: which image is more colorful
A. The first image
B. The second image
GT: B

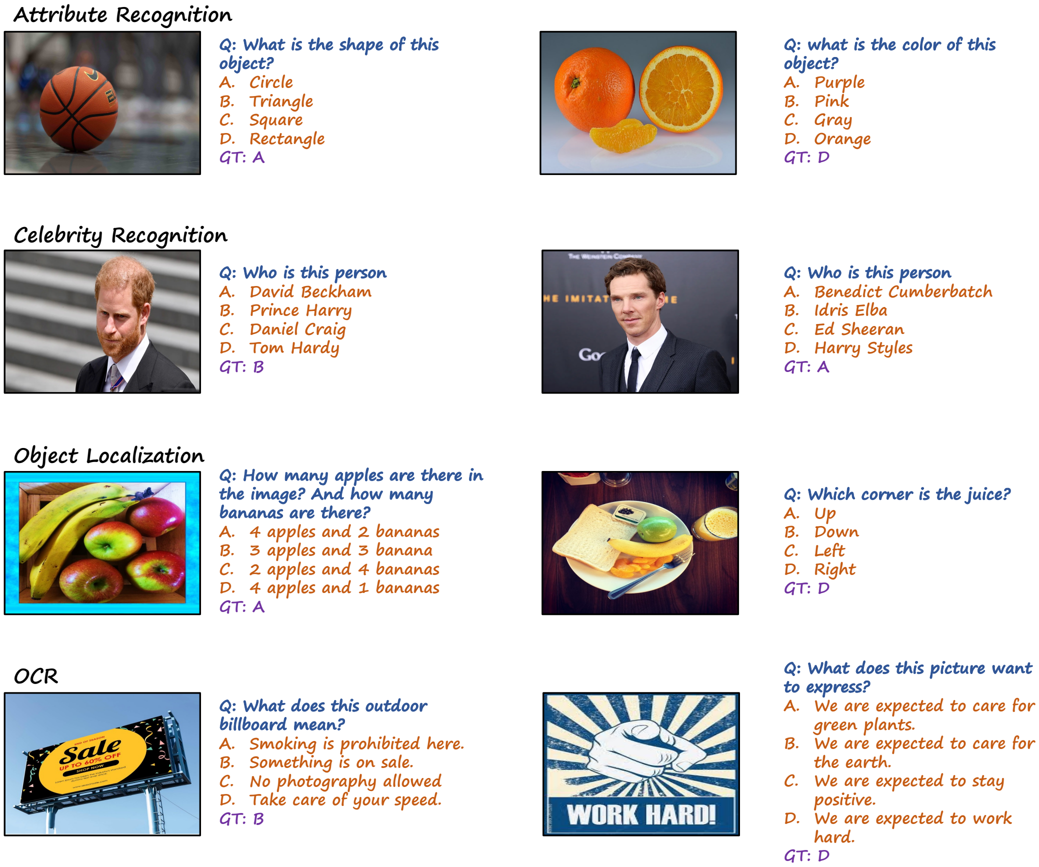
Figure 11: **Coarse Perception: Data samples.**

Coarse Perception

1. **Image Style**: Determine which type of image it belongs to, such as photos, paintings, CT scans, etc.
2. **Image Scene**: Determine which environment is shown in the image, such as indoors, outdoors, forest, city, mountains, waterfront, sunny day, rainy day, etc.
3. **Image Emotion**: Determine which subjective emotion is conveyed by the overall image, such as cold, cheerful, sad, or oppressive.
4. **Image Quality**: Determine the objective quality of the image, such as whether it is blurry, bright or dark, contrast, etc.

5. **Image Topic**: Determine what the subject of the image is, such as scenery, portrait, close-up of an object, text, etc.

In Figure 11, we visualize data samples belonging to the **Coarse Perception** capability.



Attribute Recognition

Q: What is the shape of this object?
A. Circle
B. Triangle
C. Square
D. Rectangle
GT: A

Q: what is the color of this object?
A. Purple
B. Pink
C. Gray
D. Orange
GT: D

Celebrity Recognition

Q: Who is this person
A. David Beckham
B. Prince Harry
C. Daniel Craig
D. Tom Hardy
GT: B

Q: Who is this person
A. Benedict Cumberbatch
B. Idris Elba
C. Ed Sheeran
D. Harry Styles
GT: A

Object Localization

Q: How many apples are there in the image? And how many bananas are there?
A. 4 apples and 2 bananas
B. 3 apples and 3 banana
C. 2 apples and 4 bananas
D. 4 apples and 1 bananas
GT: A

Q: Which corner is the juice?
A. Up
B. Down
C. Left
D. Right
GT: D

OCR

Q: What does this outdoor billboard mean?
A. Smoking is prohibited here.
B. Something is on sale.
C. No photography allowed
D. Take care of your speed.
GT: B

Q: What does this picture want to express?
A. We are expected to care for green plants.
B. We are expected to care for the earth.
C. We are expected to stay positive.
D. We are expected to work hard.
GT: D

Figure 12: **Fine-grained Perception (single-instance): Data samples.**

Fine-grained Perception (single-instance)

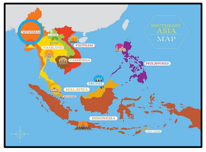
1. **Object Localization**: For a single object, determine its position in the image (such as top, bottom, etc.), its absolute coordinates in the image, count the number of objects, and the orientation of the object.
2. **Attribute Recognition**: Recognition of texture, shape, appearance characteristics, emotions, category.
3. **Celebrity Recognition**: Recognition of celebrities, landmarks, and well-known objects.
4. **OCR**: Recognition of text, formula, and sheet in the image.

In Figure 12, we visualize data samples belonging to the **Fine-grained Perception (single-instance)** capability.

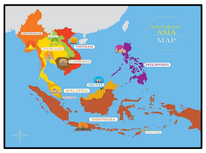
Fine-grained Perception (cross-instance)

1. **Spatial Relationship**: Determine the relative position between objects in image.
2. **Attribute Comparison**: Compare attributes of different objects in image, such as shape, color, etc.
3. **Action Recognition**: Recognizing human actions, including pose motion, human-object interaction, and human-human interaction.

Spatial Relationship

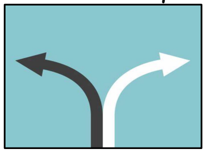


Q: Which country is north of the country circled in blue?
A. Laos
B. Thailand
C. China
D. Indonesia
GT: C

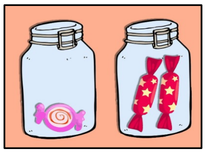


Which country is the southernmost of all the countries shown in the picture?
A. Australia
B. Indonesia
C. China
D. New Zealand
GT: B

Attribute Comparison



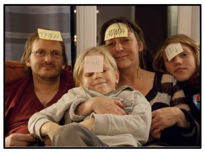
Q: Are the two arrows in the same direction in the picture?
A. Same
B. Not the same
C. Can't judge
GT: B



Q: Are the candies in the two jars in the picture the same color?
A. Same
B. Not the same
C. Can't judge
GT: B

Action Recognition

Q: What kind of human behavior does this picture describe?
A. A man with a solemn expression, XXX driving.
B. A man is practicing his skateboarding XXX skills.
C. A group of XXX breather from work.
D. A family is XXX clothing.
GT: A



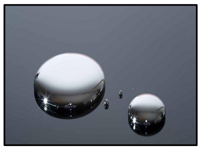
Q: What kind of human behavior does this picture describe?
A. This is a XXX smiles on their faces
B. A man is XXX his breathing and inner thoughts.
C. A musician XXX a classical piece.
D. A family is XXX together.
GT: A

Figure 13: **Fine-grained Perception (cross-instance): Data samples. XXX** indicates omitted contents which are less relevant to the question.

In Figure 13, we visualize data samples belonging to the **Fine-grained Perception (cross-instance)** capability.

Physical Property Reasoning

Q: The object shown in this figure:
A. Is the hardest naturally occurring substance on Earth.
B. Conducts electricity well at room temperature.
C. Is typically found in igneous rocks like basalt and granite.
D. Has a low melting point compared to other minerals.
GT: A



Q: The object shown in this figure:
A. Is one kind of metal that is liquid at the room temperature.
B. Can be easily dissolved in water.
C. Has a low boiling point compared to other metals.
D. Is attracted to magnets.
GT: A

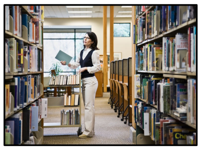
Function Reasoning

Q: What's the function of the demonstrated object?
A. Cut vegetables
B. stir
C. Water purification
D. Boiling water
GT: B

Q: What's the function of the demonstrated object?
A. Separating
B. Clamping
C. drill
D. incise
GT: A

Identity Reasoning

Q: What's the profession of the people in this picture?
A. Librarian
B. radio host
C. gardener
D. lawyer
GT: C



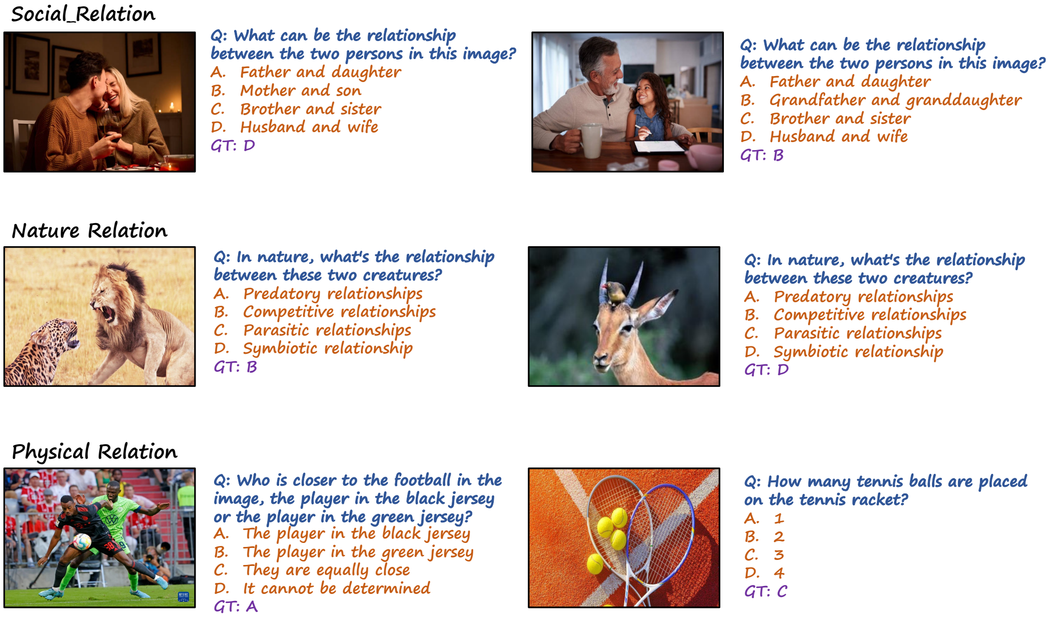
Q: What's the profession of the people in this picture?
A. Librarian
B. accountant
C. radio host
D. gardener
E. lawyer
GT: A

Figure 14: **Attribute Reasoning: Data samples.**

Attribute Reasoning

1. **Physical Property Reasoning**: Predict the physical property of an object. Examples: he physical property of concentrated sulfuric acid is that it is volatile, the physical property of water is its fluidity, etc.
2. **Function Reasoning**: Predict the function of an object. Examples: the function of a broom is to sweep the floor, the function of a spatula is to cook, the function of a pen is to write, etc.
3. **Identity Reasoning**: Predict the identity of a person. Example: by observing a person's clothing and appearance, one may infer his / her occupation.

In Figure 14, we visualize data samples belonging to the **Attribute Reasoning** capability.



Q: What can be the relationship between the two persons in this image?
A. Father and daughter
B. Mother and son
C. Brother and sister
D. Husband and wife
GT: D

Q: What can be the relationship between the two persons in this image?
A. Father and daughter
B. Grandfather and granddaughter
C. Brother and sister
D. Husband and wife
GT: B

Q: In nature, what's the relationship between these two creatures?
A. Predatory relationships
B. Competitive relationships
C. Parasitic relationships
D. Symbiotic relationship
GT: B

Q: In nature, what's the relationship between these two creatures?
A. Predatory relationships
B. Competitive relationships
C. Parasitic relationships
D. Symbiotic relationship
GT: D

Q: Who is closer to the football in the image, the player in the black jersey or the player in the green jersey?
A. The player in the black jersey
B. The player in the green jersey
C. They are equally close
D. It cannot be determined
GT: A

Q: How many tennis balls are placed on the tennis racket?
A. 1
B. 2
C. 3
D. 4
GT: C

Figure 15: **Relation Reasoning: Data samples.**

Relation Reasoning

1. **Social Relation**: Relations in human society or relations defined from the human perspective. Examples: Inter-person relations, such as father and son, husband and wife, friend, hostile, etc.
2. **Physical Relation**: All relationships that exist in the physical world, 3D spatial relationships and the connections between objects are.
3. **Nature Relation**: Other abstract relationships that exist in nature. Examples: predation, symbiosis, coexistence, etc.

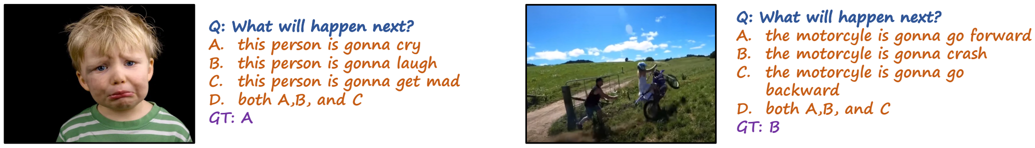
In Figure 15, we visualize data samples belonging to the **Relation Reasoning** capability.

Logic Reasoning

1. **Structuralized Image-Text Understanding**: Structured understanding of images and text, including parsing the content of charts (such as the trends of multiple bars in a bar chart), understanding the code in an image, etc.
2. **Future Prediction**: Predict what will happen in the future. Examples: if it is thundering in the sky now, it can be predicted that it will rain soon (physical phenomenon); if someone raises their fist, it means they are going to hit someone (event occurrence); if someone's face becomes serious, it means they are going to get angry (emotional change).

In Figure 16, we visualize data samples belonging to the **Logic Reasoning** capability.

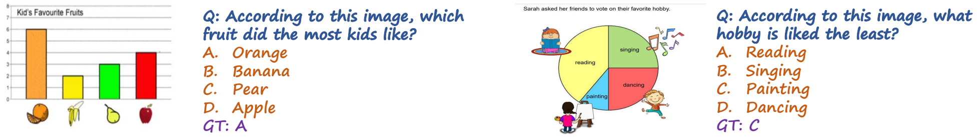
Future Prediction



Q: What will happen next?
A. this person is gonna cry
B. this person is gonna laugh
C. this person is gonna get mad
D. both A,B, and C
GT: A

Q: What will happen next?
A. the motorcyle is gonna go forward
B. the motorcyle is gonna crash
C. the motorcyle is gonna go backward
D. both A,B, and C
GT: B

Structuralized Image-text Understanding



Q: According to this image, which fruit did the most kids like?
A. Orange
B. Banana
C. Pear
D. Apple
GT: A

Q: According to this image, what hobby is liked the least?
A. Reading
B. Singing
C. Painting
D. Dancing
GT: C

Figure 16: **Logic Reasoning: Data samples.**

A.2 Data Sources of MMBench

Just as we introduce in Section 3.2 of the main paper, MMBench is mainly collected from the Internnet (80%) and the validation set of some public datasets (20%). Table 5 lists all these sources for images, questions and choices in MMBench.

Table 5: **The source of** (Q, C, I, A) **in MMBench . Customize** means all of question, choices and answer are constructed by us. **Customize & selection** implies that these components are either constructed by us or selected from the original dataset.

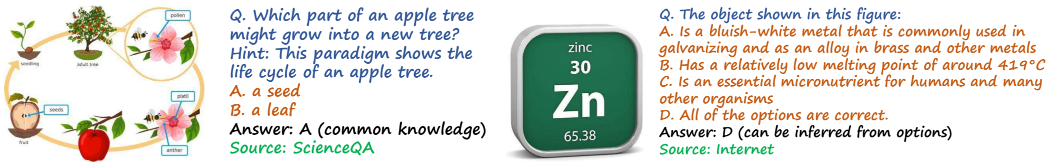
Image Source	Problem Source	Number	Ratio
ARAS [15]	customize & selection	76	2.4%
CLEVR [24]	customize & selection	14	0.4%
COCO [9]	customize & selection	179	5.6%
KonIQ-10k [21]	customize & selection	32	1.0%
LLaVA [33]	customize	19	0.6%
PISC [29]	customize & selection	15	0.5%
Places [54]	customize & selection	59	1.8%
ScienceQA [34]	customize & selection	156	4.8%
ShapeWorld [25]	customize & selection	20	0.6%
TextVQA [42]	customize & selection	18	0.6%
VSR [31]	customize & selection	19	0.6%
W3C School [1]	customize	20	0.6%
Internet	customize	2590	80.5%

B More Details on MMBench Construction

In this section we provide more qualitative results on the quality control paradigm we adopted to construct MMBench, as well as the prompt we used for MMBench-CN translation.

***'Text-only'* question filtering.** To filter out the '**text-only**' questions (which can be answered correctly with text-only inputs by LLMs) from MMBench. We apply three state-of-the-art LLMs, including GPT-4 [37], Gemini-Pro [44], and Qwen-Max [5] to infer the questions with text-only inputs under CircularEval. If more than two LLMs answer the question correctly, the question will be manually checked and removed if it is unqualified. In Figure 17(a), we visualize some unqualified questions filtered out by this approach.

***'Wrong'* question filtering.** During preliminary study, we also notice that some data samples in MMBench might be *wrong*, due to ambiguous questions or options, repeated options, or incorrect answers. To filter out these wrong samples, we infer MMBench questions with three proprietary VLMs (GPT-4v, Gemini-Pro-V, Qwen-VL-Max) and two opensource VLMs (InternLM-XComposer2



Q. Which part of an apple tree might grow into a new tree? Hint: This paradigm shows the life cycle of an apple tree.
A. a seed
B. a leaf
Answer: A (common knowledge)
Source: ScienceQA

Q. The object shown in this figure:
A. Is a bluish-white metal that is commonly used in galvanizing and as an alloy in brass and other metals
B. Has a relatively low melting point of around 419°C
C. Is an essential micronutrient for humans and many other organisms
D. All of the options are correct.
Answer: D (can be inferred from options)
Source: Internet

(a). Text-Only questions filtered out

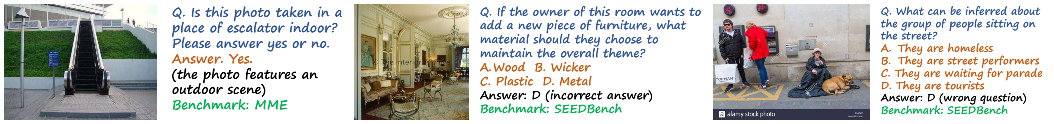
Q. What's the function of the demonstrated object?
A. Cooking
B. Cook Soup
C. Fry
D. Steam
Answer: A (ambiguous options)
Source: Internet

Q. What's the function of the demonstrated object?
A. Celebrate someone's birthday
B. Celebrating a wedding
C. A sanitary facility used for excretion
D. Offering a variety of drink
Answer: A (ambiguous options)
Source: COCO

(b). Wrong questions filtered out

Figure 17: **Unqualified samples filtered out in MMBench.**

and LLaVA-v1.5-13B). If no VLM can answer a question correctly under CircularEval, the question will then be manually checked. In Figure 17(b), we visualize wrong samples filtered out by the approach.



Q. Is this photo taken in a place of escalator indoor? Please answer yes or no.
Answer. Yes.
(the photo features an outdoor scene)
Benchmark: MME

Q. If the owner of this room wants to add a new piece of furniture, what material should they choose to maintain the overall theme?
A.Wood B. Wicker
C. Plastic D. Metal
Answer: D (incorrect answer)
Benchmark: SEEDBench

Q. What can be inferred about the group of people sitting on the street?
A. They are homeless
B. They are street performers
C. They are waiting for parade
D. They are tourists
Answer: D (wrong question)
Benchmark: SEEDBench

Figure 18: **Unqualified samples in other benchmarks can also be detected by our quality control paradigms.**

The Universality of the Quality Control Paradigm. The quality control paradigm adopted by MMBench is general and can also be applied to other benchmarks to improve the quality. To support this claim, we apply the quality control paradigm to other popular multimodal evaluation benchmarks (like MME [17] and SEEDBench [27]) and try to detect the low-quality samples. We find that our quality control paradigm can also successfully detect and filter out unqualified samples from these benchmarks. Some detected samples are visualized in Figure 18.

MMBench-CN Translation. In Figure 19, we provide the prompt we adopted for MMBench-CN translation, which include instructions and several in-context examples. All translations generated by GPT-4 will be further manually verfied to ensure the correctness.

C More Details on LLM-based Choice Extraction

Failure Cases of Heuristic Matching. In Figure 20, we display some failure cases of heuristic matching of the state-of-the-art VLM GPT-4v. Basically, such failure may occur when the VLM: i) rejects or is not capable to answer the given question; ii) answers the question in different words rather than the correct choice; iii) provides an answer with multiple choice labels (A, B, C, *etc.*) included.

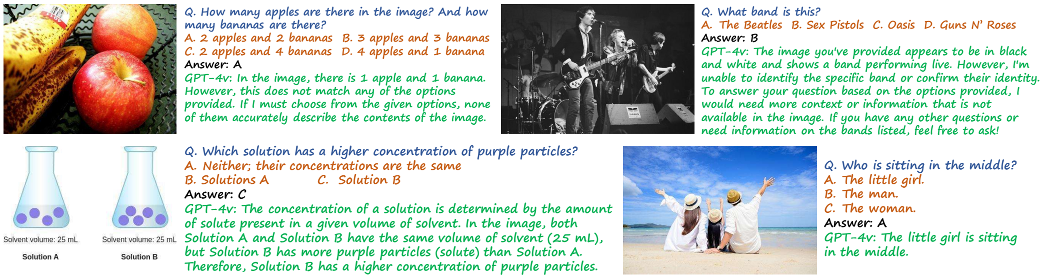
The prompt for LLM-based Choice Extraction. In Figure 21, we provide the prompt we adopted for LLM-based choice extraction. In-context examples are included to improve the instruction-following capability of the LLM adopted.

Performance Evaluated with Other Choice Extractors. In Table 6, we list the MMBench-`dev` performance obtained with different choice extractors, including GPT-4 (0125), GPT-3.5-Turbo (0613 and 0125), and InternLM2-7B [45]. VLMs with high success rate (>99%) in heuristic matching are skipped. From the table, we see that adopting different choice extractors will not lead to significant different evaluation results. VisualGLM displays the largest range across all choice extractors, which is around 1.4%. For top-performing proprietary VLMs (GPT-4v, Gemini-Pro-V, *etc.*), the gap is at most 0.3%.

B.1 MMBench-CN Translation

你是一个翻译助手，你的任务是帮我把下面的英文题目及选项翻译成中文，并保持完全一样的含义。你仅需要翻译文本中的英文内容，不需要翻译其他语言的内容，请只翻译给定内容，不要丢失/ 修改/ 添加内容。对于文本中的专有名词，符号，代码，或是人名等，请依然保持英文，不需要翻译。我会以“json” 格式给出题目及选项的内容，你需要把翻译后的中文内容以“json” 格式返回给我。
例1：
英文:
{"Q": "Which of the following was part of the role of a deaconess? ", "A": "Ministering to the sick", "B": "Preparing women for baptism", "C": "Praying for the suffering"}
中文:
{"Q": "以下哪项是女执事的职责之一？ ", "A": "照顾病人", "B": "为女性准备洗礼", "C": "为受苦的人祷告"}
例2：
英文:
{"Q": "Which can be the associated text with this image posted on twitter? ", "A": "Located in Bome County, Nyingchi City, Tibet of China, the Yigong Iron Mountain is always surrounded by clouds and mist during summer.", "B": "夏天所有季节中最闪耀的季节阳光明媚，万物清明泰山向人们展现的初夏之景处处充满着诗情画意", "C": "Giant logs and stripped trees on Rialto Beach in the Olympic National Park. #beach #wawx #blackandwhite @yourtake", "D": "Madison Falls in Olympic National Park, WA [OC] [3024x4032] #nature"}
中文:
{"Q": "与这张推特上图片配套的推文是什么？ ", "A": "坐落在中国西藏自治区林芝市波密县的易贡铁山，在夏季总是被云雾环绕。", "B": "夏天所有季节中最闪耀的季节阳光明媚，万物清明泰山向人们展现的初夏之景处处充满着诗情画意", "C": "奥林匹克国家Rialto 沙滩上的巨木与被剥皮的树木。#beach #wawx #blackandwhite @yourtake", "D": "Madison 瀑布，奥林匹克国家公园，WA [OC] [3024x4032] #nature"}
请翻译：
英文:
{The English question presented in the json format}
中文:

Figure 19: **An example prompt of Chinese single choice with reasoning.**



Q. How many apples are there in the image? And how many bananas are there?
A. 2 apples and 2 bananas B. 3 apples and 3 bananas
C. 2 apples and 4 bananas D. 4 apples and 1 banana
Answer: A
GPT-4v: In the image, there is 1 apple and 1 banana. However, this does not match any of the options provided. If I must choose from the given options, none of them accurately describe the contents of the image.

Q. What band is this?
A. The Beatles B. Sex Pistols C. Oasis D. Guns N' Roses
Answer: B
GPT-4v: The image you've provided appears to be in black and white and shows a band performing live. However, I'm unable to identify the specific band or confirm their identity. To answer your question based on the options provided, I would need more context or information that is not available in the image. If you have any other questions or need information on the bands listed, feel free to ask!

Q. Which solution has a higher concentration of purple particles?
A. Neither; their concentrations are the same
B. Solutions A C. Solution B
Answer: C
GPT-4v: The concentration of a solution is determined by the amount of solute present in a given volume of solvent. In the image, both Solution A and Solution B have the same volume of solvent (25 mL), but Solution B has more purple particles (solute) than Solution A. Therefore, Solution B has a higher concentration of purple particles.

Q. Who is sitting in the middle?
A. The little girl.
B. The man.
C. The woman.
Answer: A
GPT-4v: The little girl is sitting in the middle.

Figure 20: **Failure cases of GPT-4v during exact matching.**

LLM-based sementic matching is generally helpful. To demonstrate that LLMs can be a general tool for semantic matching, we also validate the LLM-involved evaluation paradigm on existing multi-modality tasks, including GQA [23], OK-VQA [35], and Text-VQA [42]. Given the ground-truth answer, we use GPT-3.5-Turbo to measure the similarity between VLM's prediction[5]. For each benchmark, we randomly select 1000 testing samples and evaluate with exact match (the traditional paradigm) and ChatGPT-based match, respectively, and list the results in Table 7. Basically, ChatGPT-based evaluation demonstrates the same trend compared to the exact-match accuracy on all tasks. On GQA, two algorithms demonstrate very close performance under ChatGPT-based

---

[5] The simlarity score is an integer in [1, 5]. 1 means completely wrong, while 5 means completely correct.

C.1 Prompt for Choice Extraction

You are an AI assistant who will help me to match an answer with several options of a single-choice question. You are provided with a question, several options, and an answer, and you need to find which option is most similar to the answer. If the meaning of all options are significantly different from the answer, output Z. You should only do the matching based exactly on the literal meaning of the options and answer. You should not perform any external inference based on your knowledge during the matching. Your should output a single uppercase character in A, B, C, D (if they are valid options), and Z.
Example 1:
Question: What is the main object in image?
Options: A. teddy bear B. rabbit C. cat D. dog
Answer: a cute teddy bear
Your output: A
Example 2:
Question: What is the main object in image?
Options: A. teddy bear B. rabbit C. cat D. dog
Answer: Spider
Your output: Z
Now it's your turn:
Question: {question}
Options: {options}
Answer: {answer}
Your output:

Figure 21: **The prompt used for choice extraction on MMBench.** The Chinese translation of this prompt is adopted for MMBench-CN choice extraction.

Table 6: **MMBench-`dev` accuracies with different choice extractors under *CircularEval*.**

VLM	Exact Matching	GPT-4-Turbo (0125)	GPT-3.5-Turbo (0613)	GPT-3.5-Turbo (0125)	InternLM2-7B
MiniGPT4-7B [56]	26.0	32.7	33.1	33.0	32.9
IDEFICS-9B-Instruct [26]	36.0	37.2	37.2	37.2	37.2
InstructBLIP-7B [11]	34.8	37.4	37.5	37.5	37.7
VisualGLM-6B [13]	19.4	36.1	37.5	37.5	36.1
MiniGPT4-13B [56]	30.7	37.5	37.8	37.8	37.6
InstructBLIP-13B [11]	36.6	40.9	41.1	41.0	41.4
Qwen-VL-Chat [6]	56.7	59.5	59.8	59.4	59.8
Qwen-VL-Plus [6]	43.7	62.9	62.6	61.9	63.2
MiniCPM-V [39]	57.6	64.8	64.7	64.6	64.7
Gemini-Pro-V [44]	70.4	70.9	70.9	70.9	70.8
GPT-4v [37]	71.8	74.3	74.6	74.6	74.6
Qwen-VL-Max [6]	72.9	76.4	76.5	76.2	76.5

evaluation. In further investigation, we find the reason is that ChatGPT succeeds in matching slightly different answers (compared to GT) generated by MiniGPT-4, while exact matching fails (examples in Table 8).

D Evaluation Settings and Results

In Section 5.2 of the main paper, we give the results of different models on the **test** split of MMBench and MMBench-CN. In this section, we will introduce the detailed evaluation setting on MMBench, and provide more evaluation results.

Table 7: **LLM-based Matching** vs **Exact Matching.** A preliminary study on VQA benchmarks. **Accuracy** is the success rate of answers being exactly matched with the groundtruth. For each sample, **GPT score** is an integer $n \in [1, 5]$, indicating the similarity between answer and groundtruth. We adopt **gpt-3.5-turbo-0613** as the matching LLM and report the average GPT score for testing samples.

Dataset	GQA [23]		OK-VQA [35]		Text-VQA [42]	
Model	Flamingo	MiniGPT-4	Flamingo	MiniGPT-4	Flamingo	MiniGPT-4
Accuracy	33.6%	22.4%	42.6%	21.9%	22.9%	9.8%
Average GPT score	2.75	2.74	2.79	1.97	1.92	1.54

Table 8: MiniGPT-4 GQA answers (different from GT) that can be matched by ChatGPT-based evaluation.

Question	**GroundTruth**	**MiniGPT-4 Answer**	**GPT Score**
On which side of the picture is the leather bag?	right	leather bag is on right side	5
Who is wearing a shirt?	girl	woman is wearing shirt	5
What is the pen made of?	plastic	pen is made of plastic##	5
What does the male person ride on?	bike	bicycle	4
Is the microwave silver and rectangular?	Yes	microwave is silver and rectangular	5
How does the silver lamp appear to be, on or off?	off	silver lamp appears to be off ##	5

D.1 Evaluation Settings

Unless stated otherwise, all results presented in this paper adhere to the conventional **zero-shot** evaluation setting. We have also attempted to assess these models with few-shot and chain-of-thought evaluations. **However, no encouraging results are observed.** Below we provide the prompt we used for evaluating a VLM under the zero-shot setting on MMBench.

D.1 Prompt Template for Zero-shot Inference.

Hint: xxx [optional]
Question: xxx
A. xxx
B. xxx
C. xxx [optional]
D. xxx [optional]
Please select the correct answer from the options above.

Figure 22: **The prompt template adopted for zero-shot inference.**

D.2 Model Settings

In Table 9, we provide details of all open-source models evaluated in MMBench, including several additional models that do not fit the space of the main article.

D.3 More Results

In this section, we give more detailed results about the performance of different models on MMBench and MMBench-CN. We present the detailed evaluation results of 30 different VLMs (some of them do not appear in the main paper due to limited space). For detailed results on each L-3 ability, see the separate sheet in the supplementary materials.

Table 9: **Details of the evaluated Open-Source VLMs.**

VLM	Language Backbone	Vision Backbone	Overall Parameters
OpenFlamingov2[4]	MPT 7B	CLIP ViT-L/14	9B
MiniGPT-4-7B[56]	Vicuna 7B	EVA-G	8B
IDEFICS-9B-Instruct[26]	LLaMA 7B	CLIP ViT-H/14	9B
VisualGLM-6B[13]	ChatGLM 6B	EVA-CLIP	7B
InstructBLIP-7B[11]	Vicuna 7B	EVA-G	8B
MiniGPT-4-13B[56]	Vicuna 13B	EVA-G	14B
PandaGPT[43]	Vicuna 13B	ImageBind ViT-H/14	14B
InstructBLIP-13B[11]	Vicuna 13B	EVA-G	14B
IDEFICS-80B-Instruct [26]	LLaMA 65B	CLIP ViT-H/14	80B
Qwen-VL-Chat[6]	Qwen 7B	ViT-G/16	10B
MiniCPM-V[39]	MiniCPM 2.4B	SigLip-400M	3B
LLaVA-v1.5-7B[32]	Vicuna 7B	CLIP ViT-L/14	7B
mPLUG-Owl2[49]	LLaMA2 7B	CLIP ViT-L/14	8B
CogVLM-Chat-17B[47]	Vicuna 7B	EVA2-CLIP-E	18B
ShareGPT4V-7B[8]	Vicuna 7B	CLIP ViT-L/14	7B
Yi-VL-6B[2]	Yi-6B	CLIP ViT-H/14	7B
LLaVA-InternLM-7B[10]	InternLM 7B	CLIP ViT-L/14	9B
ShareGPT4V-13B[8]	Vicuna 13B	CLIP ViT-L/14	13B
LLaVA-v1.5-13B[32]	Vicuna 13B	CLIP ViT-L/14	13B
Yi-VL-34B[2]	Yi 34B	CLIP ViT-H/14	35B
OmniLMM-12B[38]	Zephyr-7B-β	EVA-02-5B	12B
Monkey-Chat[30]	Qwen 7B	ViT BigG	10B
InternLM-XComposer[52]	InternLM-7B	EVA-G	9B
LLaVA-InternLM2-7B[10]	InternLM2-7B	CLIP ViT-L/14	9B
LLaVA-InternLM2-20B[10]	InternLM2-20B	CLIP ViT-L/14	23B
InternLM-XComposer2[12]	InternLM2-7B	CLIP ViT-L/14	9B

Table 10: **CircularEval results on MMBench-`dev` set (L-2 abilities).** Open-source models tagged with * incorporate in-house data in model training.

Model	Overall	CP	FP-S	FP-C	AR	LR	RR
OpenSource VLMs							
OpenFlamingo v2 [4]	2.6%	0.8%	4.5%	1.1%	5.5%	0.0%	3.4%
MiniGPT4-7B [56]	32.7%	38.4%	39.1%	20.7%	49.4%	10.5%	22.4%
VisualGLM-6B [13]	36.1%	40.3%	43.3%	19.6%	49.4%	16.9%	33.9%
IDEFICS-9B-Instruct [26]	37.2%	50.6%	37.7%	30.2%	51.8%	4.8%	25.3%
InstructBLIP-7B [11]	37.4%	46.4%	47.1%	23.5%	51.2%	8.1%	24.7%
MiniGPT4-13B [56]	37.5%	44.2%	48.4%	16.8%	57.3%	6.5%	30.5%
InstructBLIP-13B [11]	40.9%	48.6%	52.2%	18.4%	56.7%	5.6%	39.7%
PandaGPT [43]	41.6%	56.1%	34.6%	34.6%	53.7%	13.7%	38.5%
IDEFICS-80B-Instruct [26]	42.3%	54.7%	48.1%	24.6%	57.3%	8.9%	34.5%
Qwen-VL-Chat* [6]	59.5%	70.7%	69.9%	49.7%	69.5%	25.0%	44.3%
CogVLM-Chat-17B [47]	62.4%	69.6%	70.6%	56.4%	67.1%	29.0%	59.2%
LLaVA-v1.5-7B [32]	62.5%	71.3%	70.6%	55.9%	70.7%	25.8%	55.7%
mPLUG-Owl2 [50]	63.5%	72.9%	70.2%	53.6%	70.7%	29.8%	60.3%
MiniCPM-V [39]	64.8%	71.0%	75.1%	52.5%	72.0%	30.6%	64.9%
Yi-VL-6B* [2]	65.6%	72.7%	73.7%	54.7%	73.2%	32.3%	65.5%
ShareGPT4V-7B [8]	66.2%	77.3%	75.1%	57.5%	68.3%	25.8%	63.8%
ShareGPT4V-13B [8]	67.0%	75.1%	77.9%	58.1%	68.9%	35.5%	61.5%
LLaVA-InternLM-7B [10]	67.0%	75.7%	72.7%	57.5%	71.3%	37.1%	66.7%
LLaVA-v1.5-13B [32]	67.2%	74.0%	75.1%	59.2%	68.9%	38.7%	66.7%
Yi-VL-34B* [2]	68.2%	75.7%	73.0%	55.9%	75.6%	39.5%	70.7%
Monkey-Chat [30]	68.8%	72.9%	79.2%	58.1%	79.3%	42.7%	62.6%
OmniLMM-12B* [38]	69.7%	75.1%	79.6%	61.5%	73.8%	37.1%	69.5%
LLaVA-InternLM2-7B [10]	71.6%	79.8%	77.2%	62.0%	74.4%	41.1%	74.1%
LLaVA-InternLM2-20B [10]	72.8%	80.1%	75.1%	68.2%	73.8%	46.0%	76.4%
InternLM-XComposer* [52]	73.9%	79.6%	81.7%	65.4%	84.8%	39.5%	72.4%
InternLM-XComposer2* [12]	79.1%	83.4%	84.4%	68.7%	83.5%	58.1%	82.8%
Proprietary VLMs							
Qwen-VL-Plus [6]	62.9%	67.1%	78.9%	53.1%	71.3%	28.2%	54.6%
Gemini-Pro-V [44]	70.9%	71.3%	81.7%	62.0%	78.7%	47.6%	70.7%
GPT-4v [37]	74.3%	78.5%	72.3%	66.5%	82.9%	67.7%	73.6%
Qwen-VL-Max [6]	76.4%	76.2%	87.2%	69.3%	78.7%	55.6%	78.7%

Table 11: **CircularEval results on MMBench-`test` set (L-2 abilities).** Open-source models tagged with * incorporate in-house data in model training.

Model	**Overall**	**CP**	**FP-S**	**FP-C**	**AR**	**LR**	**RR**
OpenSource VLMs							
OpenFlamingo v2 [4]	2.3%	1.1%	3.5%	1.5%	5.3%	0.0%	2.7%
MiniGPT4-7B [56]	30.5%	37.0%	31.8%	17.2%	49.8%	9.2%	25.6%
IDEFICS-9B-Instruct [26]	35.2%	48.3%	31.3%	29.6%	47.8%	11.4%	25.2%
VisualGLM-6B [13]	35.4%	40.2%	38.5%	26.2%	47.8%	19.6%	29.5%
InstructBLIP-7B [11]	38.3%	46.7%	39.0%	31.8%	55.5%	8.7%	31.0%
MiniGPT4-13B [56]	38.8%	44.6%	42.9%	23.2%	64.9%	8.2%	32.9%
PandaGPT [43]	39.7%	51.9%	29.5%	27.3%	62.0%	19.0%	38.0%
InstructBLIP-13B [11]	39.8%	47.2%	42.9%	21.0%	60.4%	12.5%	38.8%
IDEFICS-80B-Instruct [26]	40.9%	54.6%	38.1%	29.6%	52.7%	16.8%	34.9%
Qwen-VL-Chat* [6]	60.9%	68.5%	67.7%	50.2%	78.0%	37.0%	45.7%
MiniCPM-V [39]	61.4%	65.6%	69.4%	51.3%	70.6%	35.3%	59.7%
LLaVA-v1.5-7B [32]	63.4%	70.0%	68.0%	57.7%	77.6%	33.2%	56.2%
mPLUG-Owl2 [50]	63.5%	68.1%	69.1%	55.8%	78.4%	37.0%	57.0%
CogVLM-Chat-17B [47]	63.6%	72.8%	66.6%	55.4%	71.4%	33.7%	62.0%
ShareGPT4V-7B [8]	64.6%	72.2%	68.7%	59.6%	72.7%	34.8%	60.5%
Yi-VL-6B* [2]	65.5%	72.8%	72.9%	56.2%	75.5%	41.3%	55.4%
LLaVA-InternLM-7B [10]	65.9%	72.6%	68.7%	57.3%	80.0%	37.5%	63.2%
ShareGPT4V-13B [8]	66.7%	75.6%	73.5%	56.9%	72.7%	37.0%	62.4%
LLaVA-v1.5-13B [32]	66.9%	73.1%	72.4%	60.3%	75.5%	35.9%	65.5%
Yi-VL-34B* [2]	68.4%	72.0%	78.0%	54.7%	81.2%	38.6%	68.2%
OmniLMM-12B* [38]	69.2%	72.0%	79.8%	61.0%	78.0%	40.2%	66.7%
Monkey-Chat [30]	69.6%	75.0%	75.4%	63.3%	82.4%	46.7%	58.9%
InternLM-XComposer* [52]	71.3%	75.7%	76.3%	60.3%	84.5%	44.6%	71.7%
LLaVA-InternLM2-7B [10]	71.6%	78.1%	75.4%	66.7%	77.6%	44.6%	70.2%
LLaVA-InternLM2-20B [10]	72.3%	78.3%	76.6%	68.2%	78.4%	46.2%	69.4%
InternLM-XComposer2* [12]	78.1%	80.4%	83.5%	73.0%	83.7%	63.6%	74.4%
Proprietary VLMs							
Qwen-VL-Plus [6]	64.6%	66.5%	79.1%	50.2%	73.9%	42.9%	57.8%
Gemini-Pro-V [44]	70.2%	70.0%	78.9%	65.9%	82.9%	46.2%	65.9%
GPT-4v [37]	74.3%	77.6%	73.8%	71.5%	85.3%	63.6%	68.6%
Qwen-VL-Max [6]	75.4%	74.8%	87.2%	67.0%	85.3%	54.9%	70.5%

Table 12: **CircularEval results on MMBench-CN-`dev` set (L-2 abilities).** Open-source models tagged with * incorporate in-house data in model training.

Model	**Overall**	**CP**	**FP-S**	**FP-C**	**AR**	**LR**	**RR**
OpenSource VLMs							
MiniGPT4-13B [56]	11.8%	14.6%	13.8%	14.0%	15.9%	3.2%	2.3%
MiniGPT4-7B [56]	11.9%	11.9%	14.5%	7.8%	19.5%	3.2%	10.9%
OpenFlamingo v2 [4]	14.3%	14.4%	14.9%	11.2%	21.3%	10.5%	12.6%
InstructBLIP-13B [11]	15.1%	16.0%	14.9%	7.8%	30.5%	4.0%	14.4%
InstructBLIP-7B [11]	18.1%	16.0%	16.6%	10.6%	38.4%	4.0%	23.6%
IDEFICS-9B-Instruct [26]	18.7%	22.7%	19.7%	7.3%	35.4%	1.6%	17.2%
IDEFICS-80B-Instruct [26]	29.2%	32.0%	27.0%	25.1%	50.0%	8.1%	26.4%
PandaGPT [43]	31.0%	40.1%	24.9%	18.4%	47.6%	12.1%	33.3%
VisualGLM-6B [13]	40.6%	45.3%	48.1%	30.7%	54.3%	8.9%	37.9%
CogVLM-Chat-17B [47]	52.9%	63.5%	56.4%	41.9%	65.9%	16.9%	50.0%
LLaVA-v1.5-7B [32]	57.0%	69.3%	59.9%	47.5%	62.8%	25.0%	54.0%
Qwen-VL-Chat* [6]	57.6%	66.6%	68.5%	43.6%	70.1%	21.8%	48.9%
mPLUG-Owl2 [50]	58.1%	68.8%	65.1%	43.0%	68.9%	29.8%	50.0%
ShareGPT4V-7B [8]	59.7%	71.8%	62.6%	48.6%	62.8%	26.6%	61.5%
OmniLMM-12B* [38]	60.6%	67.7%	69.9%	48.0%	70.1%	25.8%	59.2%
ShareGPT4V-13B [8]	62.4%	72.9%	67.1%	55.3%	66.5%	34.7%	55.7%
LLaVA-v1.5-13B [32]	62.5%	71.8%	65.7%	57.0%	67.1%	33.1%	59.8%
MiniCPM-V [39]	63.0%	68.2%	75.1%	53.1%	72.0%	25.8%	60.3%
LLaVA-InternLM-7B [10]	63.0%	72.4%	68.2%	50.3%	68.9%	35.5%	62.1%
Monkey-Chat [30]	65.1%	73.8%	74.4%	50.3%	77.4%	37.9%	54.6%
Yi-VL-6B* [2]	65.3%	72.4%	73.0%	53.1%	70.7%	33.9%	67.8%
Yi-VL-34B* [2]	67.0%	73.8%	73.0%	52.5%	72.6%	40.3%	71.8%
LLaVA-InternLM2-7B [10]	70.0%	81.5%	72.3%	59.2%	73.8%	34.7%	74.7%
InternLM-XComposer* [52]	71.3%	76.5%	77.5%	63.7%	81.7%	37.9%	71.8%
LLaVA-InternLM2-20B [10]	71.7%	77.9%	74.4%	68.7%	75.6%	43.5%	74.1%
InternLM-XComposer2* [12]	77.2%	83.4%	84.1%	64.2%	84.1%	54.8%	75.9%
Proprietary VLMs							
Qwen-VL-Plus [6]	67.5%	68.8%	83.0%	54.2%	75.6%	38.7%	65.5%
Gemini-Pro-V [44]	69.3%	72.4%	78.5%	63.1%	78.7%	40.3%	65.5%
GPT-4v [37]	73.3%	76.5%	71.6%	67.0%	82.3%	63.7%	74.1%
Qwen-VL-Max [6]	75.9%	73.8%	85.8%	71.5%	81.7%	55.6%	77.0%

Table 13: **CircularEval results on MMBench-CN-`test` set (L-2 abilities).** Open-source models tagged with * incorporate in-house data in model training.

Model	Overall	CP	FP-S	FP-C	AR	LR	RR
OpenSource VLMs							
MiniGPT4-7B [56]	10.8%	9.4%	11.8%	5.6%	24.5%	4.9%	8.5%
MiniGPT4-13B [56]	13.2%	16.3%	13.5%	9.0%	27.3%	3.8%	4.3%
OpenFlamingo v2 [4]	13.3%	16.5%	10.2%	9.0%	18.8%	11.4%	12.4%
InstructBLIP-13B [11]	13.7%	13.7%	14.6%	6.4%	26.5%	4.3%	14.3%
InstructBLIP-7B [11]	18.1%	15.7%	18.6%	9.4%	31.4%	8.7%	25.2%
IDEFICS-9B-Instruct [26]	19.6%	22.4%	17.4%	7.1%	35.9%	6.0%	24.4%
IDEFICS-80B-Instruct [26]	28.8%	33.0%	26.9%	25.1%	41.2%	13.6%	26.0%
PandaGPT [43]	29.6%	40.4%	20.0%	12.0%	49.8%	13.0%	34.1%
VisualGLM-6B [13]	38.1%	44.8%	39.4%	22.8%	55.5%	18.5%	34.9%
CogVLM-Chat-17B [47]	54.0%	66.1%	49.7%	47.6%	67.8%	26.1%	49.6%
LLaVA-v1.5-7B [32]	56.9%	65.2%	53.6%	52.1%	75.5%	31.0%	50.8%
Qwen-VL-Chat* [6]	57.5%	63.0%	64.5%	41.6%	74.7%	35.9%	50.0%
mPLUG-Owl2 [50]	58.0%	64.4%	57.1%	50.2%	75.1%	31.5%	56.6%
ShareGPT4V-7B [8]	58.3%	67.2%	58.2%	51.3%	72.7%	28.3%	54.7%
MiniCPM-V [39]	59.6%	64.8%	66.6%	52.8%	69.0%	33.2%	54.3%
OmniLMM-12B* [38]	60.8%	64.8%	66.4%	53.9%	74.7%	30.4%	58.9%
LLaVA-v1.5-13B [32]	62.2%	68.3%	61.5%	56.9%	73.5%	35.9%	64.3%
ShareGPT4V-13B [8]	62.7%	69.6%	63.6%	56.2%	74.7%	36.4%	60.9%
Yi-VL-6B* [2]	63.5%	68.7%	71.7%	52.4%	74.7%	39.7%	56.6%
LLaVA-InternLM-7B [10]	64.1%	70.7%	63.8%	55.8%	75.5%	39.7%	65.5%
Monkey-Chat [30]	65.0%	71.5%	68.9%	52.1%	80.0%	46.7%	57.4%
Yi-VL-34B* [2]	66.2%	69.6%	75.6%	56.2%	80.0%	37.0%	61.2%
InternLM-XComposer* [52]	69.2%	74.8%	71.7%	58.1%	80.8%	39.1%	75.6%
LLaVA-InternLM2-7B [10]	69.9%	75.4%	72.9%	63.7%	81.2%	42.4%	68.6%
LLaVA-InternLM2-20B [10]	70.3%	75.6%	73.5%	67.4%	75.1%	46.2%	69.4%
InternLM-XComposer2* [12]	77.1%	80.4%	82.8%	71.2%	88.2%	55.4%	72.1%
Proprietary VLMs							
Qwen-VL-Plus [6]	67.9%	69.6%	78.4%	60.3%	75.1%	48.9%	61.2%
Gemini-Pro-V [44]	69.2%	68.1%	77.3%	64.0%	80.4%	45.7%	69.8%
GPT-4v [37]	72.1%	75.0%	70.1%	70.0%	82.4%	60.9%	69.4%
Qwen-VL-Max [6]	73.6%	74.4%	82.6%	69.3%	79.2%	55.4%	69.0%

References

[1] W3c school. In *https://www.w3schools.com/*, 2023. 17

[2] 01-ai. Yi-vl. https://huggingface.co/01-ai/Yi-VL-34B, 2023. 4, 8, 10, 22, 23, 24, 25, 26

[3] Harsh Agrawal, Karan Desai, Yufei Wang, Xinlei Chen, Rishabh Jain, Mark Johnson, Dhruv Batra, Devi Parikh, Stefan Lee, and Peter Anderson. Nocaps: Novel object captioning at scale. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8948–8957, 2019. 3

[4] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736, 2022. 4, 7, 8, 10, 22, 23, 24, 25, 26

[5] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023. 17

[6] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 2023. 4, 8, 10, 20, 22, 23, 24, 25, 26

[7] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020. 4

[8] Lin Chen, Jisong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. Sharegpt4v: Improving large multi-modal models with better captions. *arXiv preprint arXiv:2311.12793*, 2023. 4, 22, 23, 24, 25, 26

[9] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015. 2, 3, 17

[10] XTuner Contributors. Xtuner: A toolkit for efficiently fine-tuning llm. `https://github.com/InternLM/xtuner`, 2023. 4, 10, 22, 23, 24, 25, 26

[11] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *arXiv preprint arXiv:2305.06500*, 2023. 4, 8, 10, 20, 22, 23, 24, 25, 26

[12] 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. Internlm-xcomposer2: Mastering free-form text-image composition and comprehension in vision-language large model. *arXiv preprint arXiv:2401.16420*, 2024. 4, 8, 9, 10, 22, 23, 24, 25, 26

[13] Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie 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. 10, 20, 22, 23, 24, 25, 26

[14] Haodong Duan, Junming Yang, Yuxuan Qiao, Xinyu Fang, Lin Chen, Yuan Liu, Xiaoyi Dong, Yuhang Zang, Pan Zhang, Jiaqi Wang, Dahua Lin, and Kai Chen. Vlmevalkit: An open-source toolkit for evaluating large multi-modality models, 2024. 1, 8

[15] Haodong Duan, Yue Zhao, Kai Chen, Yuanjun Xiong, and Dahua Lin. Mitigating representation bias in action recognition: Algorithms and benchmarks, 2022. 17

[16] Jerry A Fodor. *The modularity of mind*. MIT press, 1983. 4

[17] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Zhenyu Qiu, Wei Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, and Rongrong Ji. Mme: A comprehensive evaluation benchmark for multimodal large language models. *ArXiv*, abs/2306.13394, 2023. 3, 18

[18] Tao Gong, Chengqi Lyu, Shilong Zhang, Yudong Wang, Miao Zheng, Qian Zhao, Kuikun Liu, Wenwei Zhang, Ping Luo, and Kai Chen. Multimodal-gpt: A vision and language model for dialogue with humans, 2023. 2

[19] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913, 2017. 2, 3

[20] Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3608–3617, 2018. 3

[21] V. Hosu, H. Lin, T. Sziranyi, and D. Saupe. Koniq-10k: An ecologically valid database for deep learning of blind image quality assessment. *IEEE Transactions on Image Processing*, 29:4041–4056, 2020. 17

[22] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. 4

[23] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709, 2019. 2, 3, 19, 21

[24] Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2901–2910, 2017. 17

[25] Alexander Kuhnle and Ann Copestake. Shapeworld-a new test methodology for multimodal language understanding. *arXiv preprint arXiv:1704.04517*, 2017. 17

[26] Hugo Laurençon, Lucile Saulnier, Léo Tronchon, Stas Bekman, Amanpreet Singh, Anton Lozhkov, Thomas Wang, Siddharth Karamcheti, Alexander M. Rush, Douwe Kiela, Matthieu Cord, and Victor Sanh. Obelics: An open web-scale filtered dataset of interleaved image-text documents, 2023. 8, 10, 20, 22, 23, 24, 25, 26

[27] Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*, 2023. 18

[28] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023. 4

[29] Junnan Li, Yongkang Wong, Qi Zhao, and Mohan S Kankanhalli. Dual-glance model for deciphering social relationships. In *Proceedings of the IEEE international conference on computer vision*, pages 2650–2659, 2017. 17

[30] Zhang Li, Biao Yang, Qiang Liu, Zhiyin Ma, Shuo Zhang, Jingxu Yang, Yabo Sun, Yuliang Liu, and Xiang Bai. Monkey: Image resolution and text label are important things for large multi-modal models. *arXiv preprint arXiv:2311.06607*, 2023. 22, 23, 24, 25, 26

[31] Fangyu Liu, Guy Edward Toh Emerson, and Nigel Collier. Visual spatial reasoning. *Transactions of the Association for Computational Linguistics*, 2023. 17

[32] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*, 2023. 4, 8, 10, 22, 23, 24, 25, 26

[33] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*, 2023. 2, 4, 9, 17

[34] Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, 35:2507–2521, 2022. 3, 17

[35] Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*, pages 3195–3204, 2019. 2, 3, 6, 19, 21

[36] Mike Oaksford and Nick Chater. *Bayesian rationality: The probabilistic approach to human reasoning*. Oxford University Press, 2007. 4

[37] OpenAI. Gpt-4 technical report. *ArXiv*, abs/2303.08774, 2023. 1, 2, 4, 5, 8, 10, 17, 20, 23, 24, 25, 26

[38] OpenBMB. Omnilmm: Large multi-modal models for strong performance and efficient deployment. `https://github.com/OpenBMB/OmniLMM`, 2023. 22, 23, 24, 25, 26

[39] OpenBMB. Minicpm: Unveiling the potential of end-side large language models, 2024. 8, 10, 20, 22, 23, 24, 25, 26

[40] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022. 4

[41] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019. 4

[42] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8317–8326, 2019. 3, 17, 19, 21

[43] Yixuan Su, Tian Lan, Huayang Li, Jialu Xu, Yan Wang, and Deng Cai. Pandagpt: One model to instruction-follow them all. *arXiv preprint arXiv:2305.16355*, 2023. 22, 23, 24, 25, 26

[44] Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023. 2, 4, 5, 8, 10, 17, 20, 23, 24, 25, 26

[45] InternLM Team. Internlm: A multilingual language model with progressively enhanced capabilities. `https://github.com/InternLM/InternLM-techreport`, 2023. 8, 9, 18

[46] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 4

[47] Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, Jiazheng Xu, Bin Xu, Juanzi Li, Yuxiao Dong, Ming Ding, and Jie Tang. Cogvlm: Visual expert for pretrained language models. *ArXiv*, abs/2311.03079, 2023. 8, 10, 22, 23, 24, 25, 26

[48] Peng Xu, Wenqi Shao, Kaipeng Zhang, Peng Gao, Shuo Liu, Meng Lei, Fanqing Meng, Siyuan Huang, Yu Jiao Qiao, and Ping Luo. Lvlm-ehub: A comprehensive evaluation benchmark for large vision-language models. 2023. 2

[49] Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, et al. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*, 2023. 2, 3, 4, 22

[50] Qinghao Ye, Haiyang Xu, Jiabo Ye, Mingshi Yan, Anwen Hu, Haowei Liu, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration. *ArXiv*, abs/2311.04257, 2023. 8, 10, 23, 24, 25, 26

[51] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014. 3

[52] Pan Zhang, Xiaoyi Dong Bin Wang, Yuhang Cao, Chao Xu, Linke Ouyang, Zhiyuan Zhao, Shuangrui Ding, Songyang Zhang, Haodong Duan, Hang Yan, et al. Internlm-xcomposer: A vision-language large model for advanced text-image comprehension and composition. *arXiv preprint arXiv:2309.15112*, 2023. 22, 23, 24, 25, 26

[53] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric. P Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena, 2023. 2, 4, 9

[54] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017. 17

[55] Luowei Zhou, Chenliang Xu, and Jason Corso. Towards automatic learning of procedures from web instructional videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018. 3

[56] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023. 2, 4, 8, 10, 20, 22, 23, 24, 25, 26