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

Yuan Liu[1,*], Haodong Duan[1,*,‡], Yuanhan Zhang[3,*], Bo Li[3,*], Songyang Zhang[1,*], Wangbo Zhao[4], Yike Yuan[5], Jiaqi Wang[1], Conghui He[1], Ziwei Liu[3,†], Kai Chen[1,†], and Dahua Lin[1,2†]

[1] Shanghai AI Laboratory [2] The Chinese University of Hong Kong
[3] S-Lab, Nanyang Technological University [4] National University of Singapore
[5] Zhejiang University
[*] Equal Contribution [‡] Project Lead [†] Corresponding Author
https://mmbench.opencompass.org.cn/home

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. MMBench has been supported in VLMEvalKit[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

[1] https://github.com/open-compass/VLMEvalKit

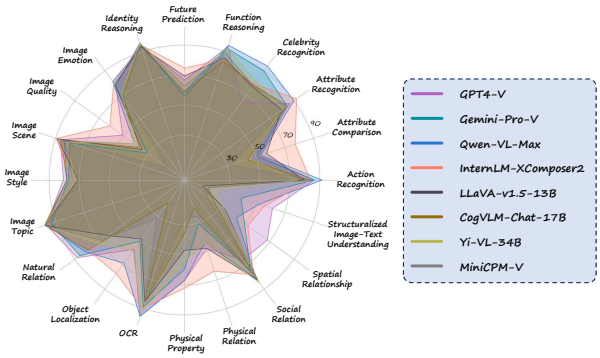

Fig. 1: Results of eight representative large vision-language models (VLMs) across the 20 ability dimensions defined in MMBench-test.

and GPT-4 [26], 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 GPT-4v [26], Gemini-Pro-V [31] and LLaVA [22], 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 [14, 22, 41] 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 [7, 15] for objective evaluation, while the second approach employs human annotators [35, 36] to perform subjective evaluations. However, it is worth noting that both approaches exhibit some inherent limitations.

A multitude of public datasets, such as VQAv2 [15], COCO Caption [7], GQA [18], and OK-VQA [24], 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 [36] and LVLM-eHub [35] 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 [39], 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, 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, small evaluation datasets can result in statistical instability. To ensure a robust evaluation, collecting more data is 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 benchmark to robustly evaluate different abilities of LVLMs. MMBench contains over 3000 multiple-choice questions covering 20 ability dimensions, such as object localization and social reasoning, for evaluating VLMs. 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. Since some existing VLMs have limited instruction-following capability and cannot directly output choice labels (A, B, C, *etc.*) for multi-choice questions, the evaluation based on exact matching may not yield accurate and reasonable conclusions. To reduce the number of false-negative samples during answer matching, we employ LLMs to match a model's prediction to candidate choices and then output the label for the matched choice. We conduct a comparison between LLM-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 [7], Nocaps [2], and Flickr30k [38] 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 [18], OK-VQA [24], VQAv2 [15], and Vizwiz [16], 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 [30] proposes questions about text shown in the image, thus involving the OCR task in question-answering. ScienceQA [23] focuses on scientific topics, requiring the model to integrate commonsense into reasoning. Youcook2 [40] 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. [36] constructed OwlEval, an evaluation set encompassing a variety of visual-related tasks, albeit of a limited size. Fu et al. [13] 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 [5, 28, 29], LLaMA [33], and Vicuna [39], recent advancements have been made in multimodal models. Flamingo [3], 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 [3] offers an open-source version of this model. BLIP-2 [20] introduces a Querying Transformer (Q-former) to bridge the modality gap between the frozen image encoder and the large language encoder. Subsequently, InstructBLIP [9] extends BLIP-2 [20] with vision-language instruction tuning, achieving superior performance. MiniGPT-4 [41] attributes the prowess of GPT-4 [26] to advanced LLMs and proposes the use of a single projection layer to align the visual representation with the language model. LLaVA [22] 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 [1,6,8,21]. During the instruction tuning, Low-Rank Adaptation (LoRA [17]) has been adopted by recent works [8, 10, 36] 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 [26], Gemini-Pro-V [31], and Qwen-VL-Max [4]. 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.

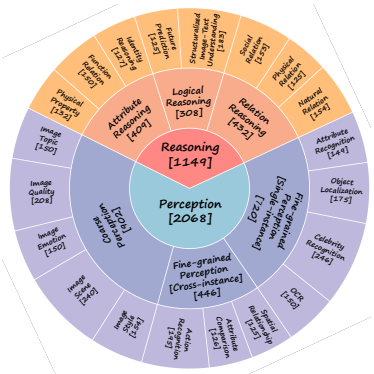

Fig. 2: Ability dimensions in MMBench. Currently, MMBench incorporates three levels of ability dimensions, encompassing 20 distinct leaf abilities.

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 [12, 25]. In pursuit of genuine

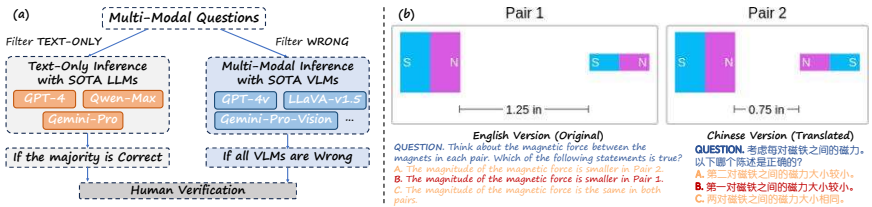

Fig. 3: The construction of MMBench. (a). The quality control strategies adopted in MMBench; (b) An illustration of questions in MMBench-CN.

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 Fig. 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 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 Fig. 3(a). We adopt 'majority voting' to detect **text-only** samples: data samples are inferred with state-of-the-art LLMs (GPT-4 [26], Gemini-Pro [31], *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**

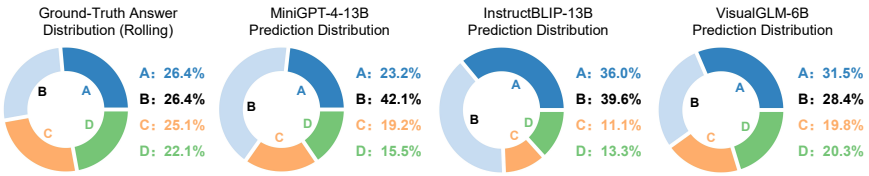

Fig. 4: The choice distribution of ground-truth answers and predictions of sample VLMs (all *CircularEval* records). Since there exist questions with only 2/3 choices in MMBench, the choice distribution of ground-truth is not exactly even.

samples, we also implement an automatic filtering mechanism. We select several state-of-the-art VLMs (including both open-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 Fig. 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 Fig. 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 [24] 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:

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,

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

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

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

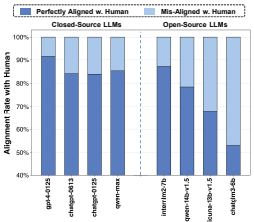

Fig. 5: Alignment rates between human and different LLMs. 'chatgpt' is 'gpt-3.5-turbo'. Open-source LLMs are 'chat' variants.

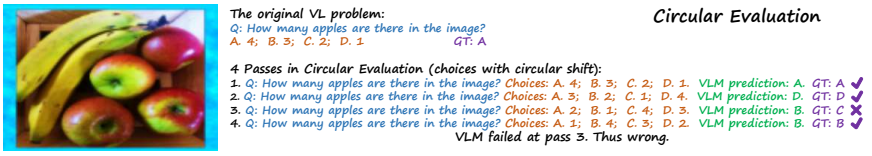

Fig. 6: CircularEval strategy. In **CircularEval**, a problem is tested multiple times with circular shifted choices and the VLM needs to succeed in all testing passes. In this example, the VLM failed in pass 3 and thus considered failed the problem.

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 [3] demonstrates top IF capability among all VLMs, while also achieving one of the worst performances on MMBench (Table 3).

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.

Fig. 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 [32] 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 ∼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 (Fig. 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 Fig. 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 [26]; (b) *Open-Source VLMs* including model variants of OpenFlamingo [3], MiniGPT4 [41], InstructBLIP [9], LLaVA [21], IDEFICS [19], CogVLM [34], Qwen-VL [4], Yi-VL [1], mPLUG-Owl [37], InternLM-XComposer [10], and MiniCPM-V [27]; (c) *Proprietary VLMs* including Qwen-VL-[Plus/Max] [4], Gemini-Pro-V [31], and GPT-4v [26]. 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.

5.2 Main Results

CircularEval *vs.* VanillaEval. 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

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%

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, *etc.*) suffer from ∼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 [10] 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 [22] (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 [39] to the more powerful InternLM2-20B [32] leads to steady improvement across all L-2 capabilities (especially significant for reasoning tasks). The scaling also holds for variants with different

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

Table 3: CircularEval results on MMBench `test` set (L-2 abilities). Abbreviations adopted: Logical Reasoning (LR), Attribute Reasoning (AR), Relation Reasoning (RR), Fine-grained Perception, X-Instance (FP-C), Fine-grained Perception, Single Instance (FP-S), Coarse Perception (CP). 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) [26]	2.9%	0.6%	1.2%	4.1%	3.7%	4.9%	7.4%
OpenSource VLMs							
OpenFlamingo v2 [3]	2.3%	1.1%	3.5%	1.5%	5.3%	0.0%	2.7%
MiniGPT4-7B [41]	30.5%	37.0%	31.8%	17.2%	49.8%	9.2%	25.6%
IDEFICS-9B-Instruct [19]	35.2%	48.3%	31.3%	29.6%	47.8%	11.4%	25.2%
VisualGLM-6B [11]	35.4%	40.2%	38.5%	26.2%	47.8%	19.6%	29.5%
InstructBLIP-7B [9]	38.3%	46.7%	39.0%	31.8%	55.5%	8.7%	31.0%
MiniGPT4-13B [41]	38.8%	44.6%	42.9%	23.2%	64.9%	8.2%	32.9%
InstructBLIP-13B [9]	39.8%	47.2%	42.9%	21.0%	60.4%	12.5%	38.8%
Qwen-VL-Chat* [4]	60.9%	68.5%	67.7%	50.2%	78.0%	37.0%	45.7%
MiniCPM-V [27]	61.4%	65.6%	69.4%	51.3%	70.6%	35.3%	59.7%
LLaVA-v1.5-7B [21]	63.4%	70.0%	68.0%	57.7%	77.6%	33.2%	56.2%
mPLUG-Owl2 [37]	63.5%	68.1%	69.1%	55.8%	78.4%	37.0%	57.0%
CogVLM-Chat-17B [34]	63.6%	72.8%	66.6%	55.4%	71.4%	33.7%	62.0%
Yi-VL-6B* [1]	65.5%	72.8%	72.9%	56.2%	75.5%	41.3%	55.4%
LLaVA-v1.5-13B [21]	66.9%	73.1%	72.4%	60.3%	75.5%	35.9%	65.5%
Yi-VL-34B* [1]	68.4%	72.0%	78.0%	54.7%	81.2%	38.6%	68.2%
LLaVA-InternLM2-20B [8]	72.3%	78.3%	76.6%	68.2%	78.4%	46.2%	69.4%
InternLM-XComposer2* [10]	78.1%	80.4%	83.5%	73.0%	83.7%	63.6%	74.4%
Proprietary VLMs							
Qwen-VL-Plus [4]	64.6%	66.5%	79.1%	50.2%	73.9%	42.9%	57.8%
Gemini-Pro-V [31]	70.2%	70.0%	78.9%	65.9%	82.9%	46.2%	65.9%
GPT-4v [26]	74.3%	77.6%	73.8%	71.5%	85.3%	63.6%	68.6%
Qwen-VL-Max [4]	75.4%	74.8%	87.2%	67.0%	85.3%	54.9%	70.5%

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. Fig. 7 compares 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

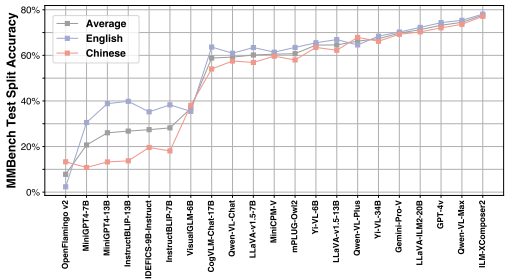

Fig. 7: The performance on the `test` split of MMBench and MMBench-CN. Models are sorted with the ascending order of average performance. ILM stands for InternLM.

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

Fig. 8: Content Moderation Cases of Proprietary VLMs.

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. The 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

Content Moderation of Proprietary VLMs. Taking an in-depth look at 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 (Fig. 8), while no obvious rejection pattern is observed for Gemini. 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 significant.

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 Fig. 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 Fig. 10. Generally, we find that all existing VLMs have the following limitations: 1. Poor at recognizing the low-level features on visual inputs,

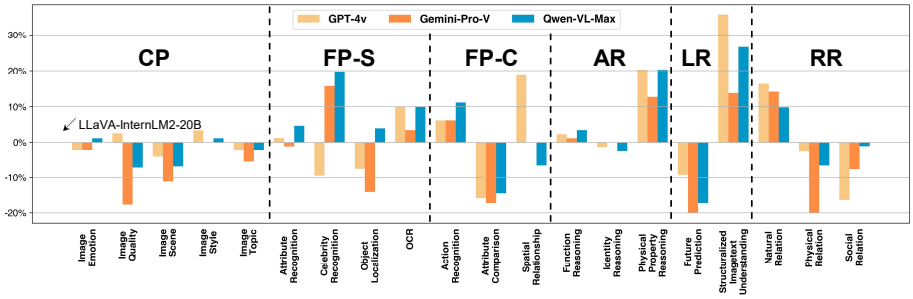

Fig. 9: Proprietary VLMs *vs.* Open-Source ones at a fine-grained level.

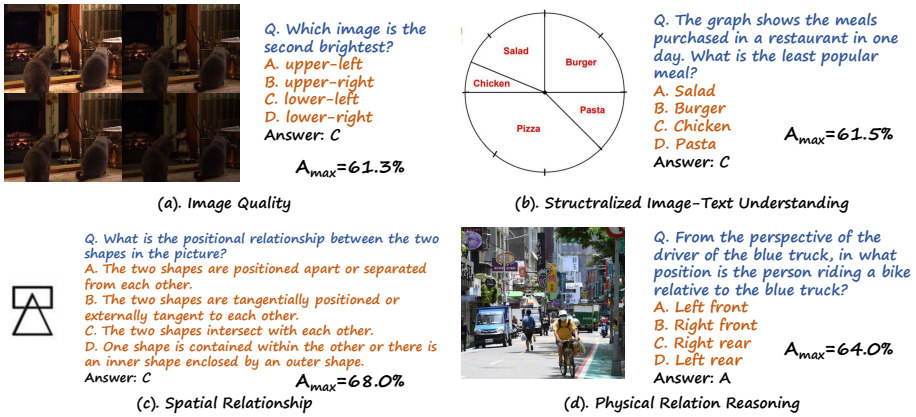

Fig. 10: Hard examples that belong to the 4 L-3 abilities with lowest A_{max}. All VLMs have made the wrong prediction for the visualized examples under CircularEval.

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 Fig. 10(b); 3. Perform badly on recognizing or reasoning about the inter-object spatial relationships, either in 2D or 3D space.

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.

Acknowledgement

This project is supported by the National Key R&D Program of China (No.2022ZD 0161600), the Shanghai Postdoctoral Excellence Program (No.2023023), China Postdoctoral Science Fund (No.2024M751559), and Shanghai Artificial lntelligence Laboratory. This project is also supported under the RIE2020 Industry Alignment Fund - Industry Collaboration Projects (IAF-ICP) Funding Initiative, as well as cash and in-kind contribution from the industry partner(s).

References

1. 01-ai: Yi-vl. https://huggingface.co/01-ai/Yi-VL-34B (2023) 5, 10, 12
2. Agrawal, H., Desai, K., Wang, Y., Chen, X., Jain, R., Johnson, M., Batra, D., Parikh, D., Lee, S., Anderson, P.: Nocaps: Novel object captioning at scale. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 8948–8957 (2019) 4
3. Alayrac, J.B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., et al.: Flamingo: a visual language model for few-shot learning. Advances in Neural Information Processing Systems **35**, 23716–23736 (2022) 4, 5, 9, 10, 12
4. Bai, J., Bai, S., Yang, S., Wang, S., Tan, S., Wang, P., Lin, J., Zhou, C., Zhou, J.: Qwen-vl: A frontier large vision-language model with versatile abilities. arXiv preprint arXiv:2308.12966 (2023) 5, 10, 12
5. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners. Advances in neural information processing systems **33**, 1877–1901 (2020) 4
6. Chen, L., Li, J., Dong, X., Zhang, P., He, C., Wang, J., Zhao, F., Lin, D.: Sharegpt4v: Improving large multi-modal models with better captions. arXiv preprint arXiv:2311.12793 (2023) 5
7. Chen, X., Fang, H., Lin, T.Y., Vedantam, R., Gupta, S., Dollár, P., Zitnick, C.L.: Microsoft coco captions: Data collection and evaluation server. arXiv preprint arXiv:1504.00325 (2015) 2, 4
8. Contributors, X.: Xtuner: A toolkit for efficiently fine-tuning llm. https://github.com/InternLM/xtuner (2023) 5, 12
9. Dai, W., Li, J., Li, D., Tiong, A.M.H., Zhao, J., Wang, W., Li, B., Fung, P., Hoi, S.: Instructblip: Towards general-purpose vision-language models with instruction tuning. arXiv preprint arXiv:2305.06500 (2023) 5, 10, 12
10. Dong, X., Zhang, P., Zang, Y., Cao, Y., Wang, B., Ouyang, L., Wei, X., Zhang, S., Duan, H., Cao, M., Zhang, W., Li, Y., Yan, H., Gao, Y., Zhang, X., Li, W., Li, J., Chen, K., He, C., Zhang, X., Qiao, Y., Lin, D., Wang, J.: Internlm-xcomposer2: Mastering free-form text-image composition and comprehension in vision-language large model. arXiv preprint arXiv:2401.16420 (2024) 5, 10, 11, 12
11. Du, Z., Qian, Y., Liu, X., Ding, M., Qiu, J., Yang, Z., Tang, J.: 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). pp. 320–335 (2022) 12
12. Fodor, J.A.: The modularity of mind. MIT press (1983) 5

13. Fu, C., Chen, P., Shen, Y., Qin, Y., Zhang, M., Lin, X., Qiu, Z., Lin, W., Yang, J., Zheng, X., Li, K., Sun, X., Ji, R.: Mme: A comprehensive evaluation benchmark for multimodal large language models. ArXiv **abs/2306.13394** (2023), https://api.semanticscholar.org/CorpusID:259243928 4
14. Gong, T., Lyu, C., Zhang, S., Wang, Y., Zheng, M., Zhao, Q., Liu, K., Zhang, W., Luo, P., Chen, K.: Multimodal-gpt: A vision and language model for dialogue with humans (2023) 2
15. Goyal, Y., Khot, T., Summers-Stay, D., Batra, D., Parikh, D.: 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. pp. 6904–6913 (2017) 2, 4
16. Gurari, D., Li, Q., Stangl, A.J., Guo, A., Lin, C., Grauman, K., Luo, J., Bigham, J.P.: Vizwiz grand challenge: Answering visual questions from blind people. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3608–3617 (2018) 4
17. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: Lora: Low-rank adaptation of large language models. arXiv preprint arXiv:2106.09685 (2021) 5
18. Hudson, D.A., Manning, C.D.: 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. pp. 6700–6709 (2019) 2, 4
19. Laurençon, H., Saulnier, L., Tronchon, L., Bekman, S., Singh, A., Lozhkov, A., Wang, T., Karamcheti, S., Rush, A.M., Kiela, D., Cord, M., Sanh, V.: Obelics: An open web-scale filtered dataset of interleaved image-text documents (2023) 10, 12
20. Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. arXiv preprint arXiv:2301.12597 (2023) 5
21. Liu, H., Li, C., Li, Y., Lee, Y.J.: Improved baselines with visual instruction tuning. arXiv preprint arXiv:2310.03744 (2023) 5, 10, 12
22. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. arXiv preprint arXiv:2304.08485 (2023) 2, 5, 11
23. Lu, P., Mishra, S., Xia, T., Qiu, L., Chang, K.W., Zhu, S.C., Tafjord, O., Clark, P., Kalyan, A.: Learn to explain: Multimodal reasoning via thought chains for science question answering. Advances in Neural Information Processing Systems **35**, 2507–2521 (2022) 4
24. Marino, K., Rastegari, M., Farhadi, A., Mottaghi, R.: Ok-vqa: A visual question answering benchmark requiring external knowledge. In: Proceedings of the IEEE/cvf conference on computer vision and pattern recognition. pp. 3195–3204 (2019) 2, 4, 7
25. Oaksford, M., Chater, N.: Bayesian rationality: The probabilistic approach to human reasoning. Oxford University Press (2007) 5
26. OpenAI: Gpt-4 technical report. ArXiv **abs/2303.08774** (2023) 2, 5, 6, 10, 12
27. OpenBMB: Minicpm: Unveiling the potential of end-side large language models (2024) 10, 12
28. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al.: Training language models to follow instructions with human feedback. Advances in Neural Information Processing Systems **35**, 27730–27744 (2022) 4
29. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al.: Language models are unsupervised multitask learners. OpenAI blog **1**(8), 9 (2019) 4

30. Singh, A., Natarajan, V., Shah, M., Jiang, Y., Chen, X., Batra, D., Parikh, D., Rohrbach, M.: Towards vqa models that can read. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8317–8326 (2019) 4
31. Team, G., Anil, R., Borgeaud, S., Wu, Y., Alayrac, J.B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A.M., Hauth, A., et al.: Gemini: a family of highly capable multimodal models. arXiv preprint arXiv:2312.11805 (2023) 2, 5, 6, 10, 12
32. Team, I.: Internlm: A multilingual language model with progressively enhanced capabilities. https://github.com/InternLM/InternLM-techreport (2023) 9, 11
33. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al.: Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971 (2023) 4
34. Wang, W., Lv, Q., Yu, W., Hong, W., Qi, J., Wang, Y., Ji, J., Yang, Z., Zhao, L., Song, X., Xu, J., Xu, B., Li, J., Dong, Y., Ding, M., Tang, J.: Cogvlm: Visual expert for pretrained language models. ArXiv **abs/2311.03079** (2023), https://api.semanticscholar.org/CorpusID:265034288 10, 12
35. Xu, P., Shao, W., Zhang, K., Gao, P., Liu, S., Lei, M., Meng, F., Huang, S., Qiao, Y.J., Luo, P.: Lvlm-ehub: A comprehensive evaluation benchmark for large vision-language models (2023) 2
36. Ye, Q., Xu, H., Xu, G., Ye, J., Yan, M., Zhou, Y., Wang, J., Hu, A., Shi, P., Shi, Y., et al.: mplug-owl: Modularization empowers large language models with multimodality. arXiv preprint arXiv:2304.14178 (2023) 2, 4, 5
37. Ye, Q., Xu, H., Ye, J., Yan, M., Hu, A., Liu, H., Qian, Q., Zhang, J., Huang, F., Zhou, J.: mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration. ArXiv **abs/2311.04257** (2023), https://api.semanticscholar.org/CorpusID:265050943 10, 12
38. Young, P., Lai, A., Hodosh, M., Hockenmaier, J.: 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) 4
39. Zheng, L., Chiang, W.L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E.P., Zhang, H., Gonzalez, J.E., Stoica, I.: Judging llm-as-a-judge with mt-bench and chatbot arena (2023) 3, 4, 11
40. Zhou, L., Xu, C., Corso, J.: Towards automatic learning of procedures from web instructional videos. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 32 (2018) 4
41. Zhu, D., Chen, J., Shen, X., Li, X., Elhoseiny, M.: Minigpt-4: Enhancing vision-language understanding with advanced large language models. arXiv preprint arXiv:2304.10592 (2023) 2, 5, 10, 12