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

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

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 Fig. 1, we visualize data samples belonging to the **Coarse Perception** capability.

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.

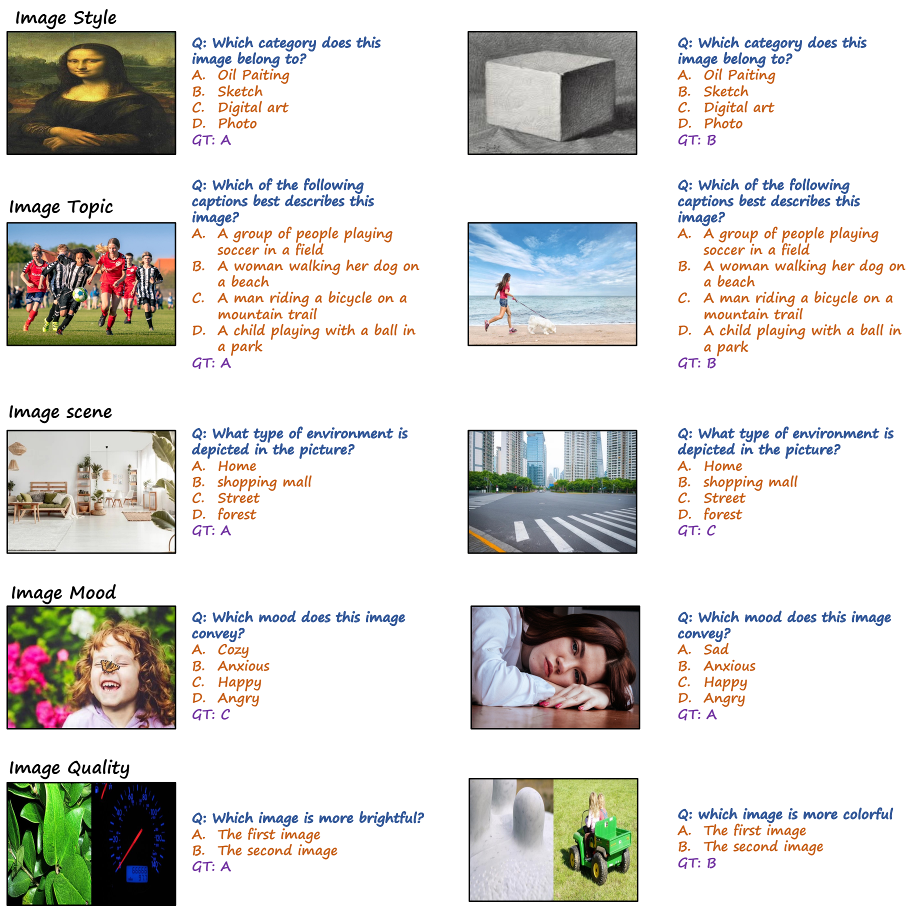

Fig. 1: Coarse Perception: Data samples.

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 Fig. 2, 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.

Fig. 2: Fine-grained Perception (single-instance): Data samples.

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.

In Fig. 3, we visualize data samples belonging to the **Fine-grained Perception (cross-instance)** capability.

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.

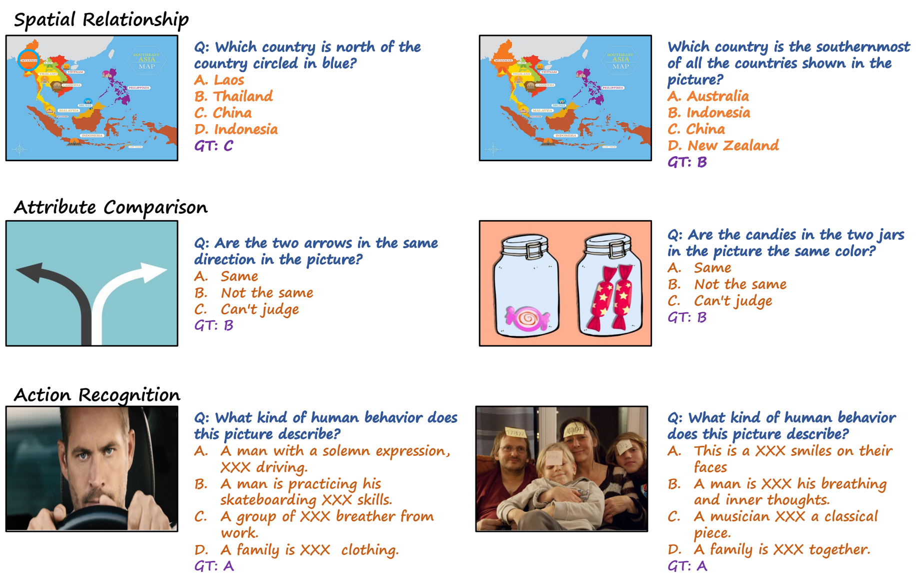

Fig. 3: Fine-grained Perception (cross-instance): Data samples. XXX indicates omitted contents which are less relevant to the question.

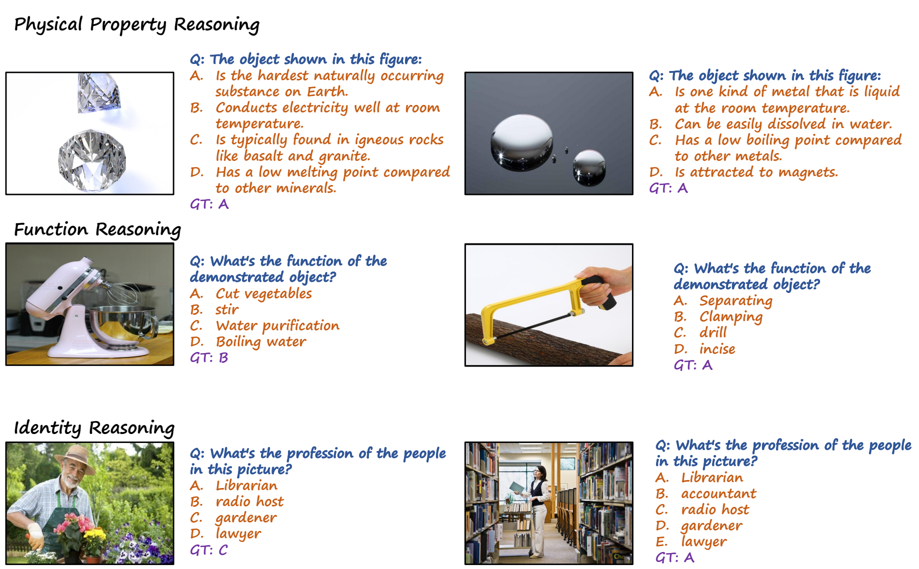

Fig. 4: Attribute Reasoning: Data samples.

In Fig. 4, we visualize data samples belonging to the **Attribute Reasoning** capability.

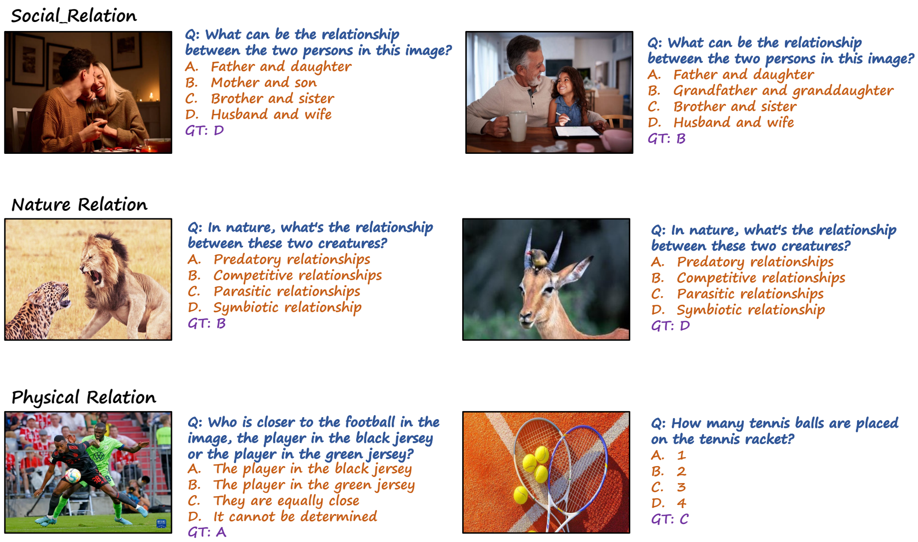

Fig. 5: Relation Reasoning: Data samples.

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 Fig. 5, 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 Fig. 6, we visualize data samples belonging to the **Logic Reasoning** capability.

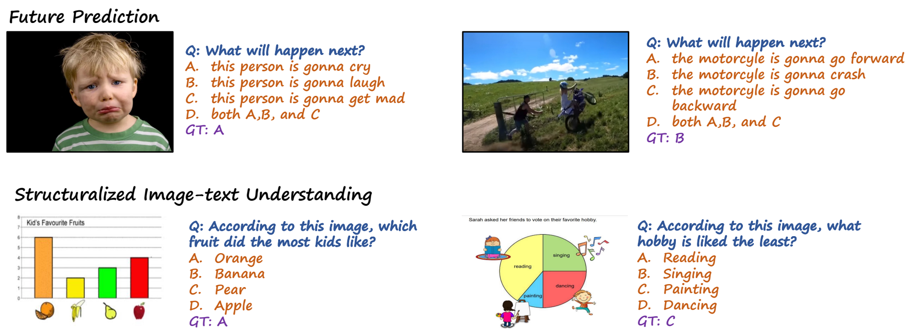

Fig. 6: Logic Reasoning: Data samples.

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 1 lists all these sources for images, questions and choices in MMBench.

Table 1: 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 [12]	customize & selection	76	2.4%
CLEVR [16]	customize & selection	14	0.4%
COCO [7]	customize & selection	179	5.6%
KonIQ-10k [14]	customize & selection	32	1.0%
LLaVA [24]	customize	19	0.6%
PISC [20]	customize & selection	15	0.5%
Places [38]	customize & selection	59	1.8%
ScienceQA [25]	customize & selection	156	4.8%
ShapeWorld [17]	customize & selection	20	0.6%
TextVQA [30]	customize & selection	18	0.6%
VSR [22]	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.

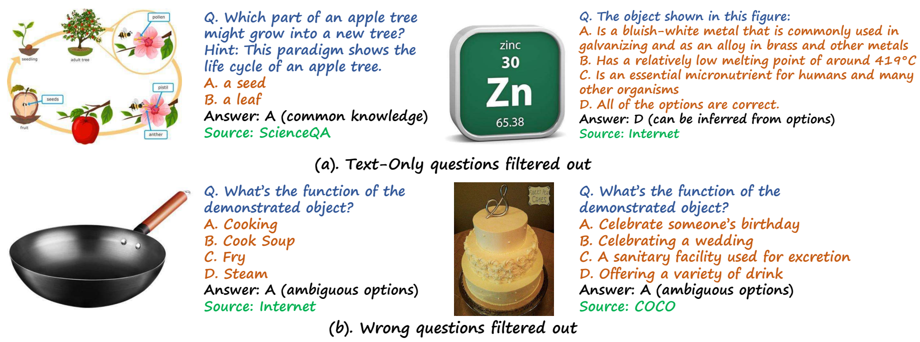

Fig. 7: Unqualified samples filtered out in MMBench.

***'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 [27], Gemini-Pro [32], and Qwen-Max [4] 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 Fig. 8(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 and LLaVA-v1.5-13B). If no VLM can answer a question correctly under CircularEval, the question will then be manually checked. In Fig. 8(b), we visualize wrong samples filtered out by the approach.

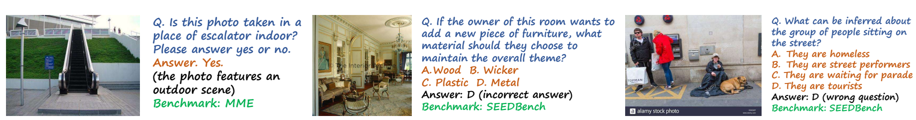

Fig. 8: Unqualified samples in other benchmarks can also be detected by our quality control paradigms.

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 [13] and SEEDBench [19]) 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 Fig. 8.

MMBench-CN Translation. In Fig. 9, 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.

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:

Fig. 9: An example prompt of Chinese single choice with reasoning.

C More Details on LLM-based Choice Extraction

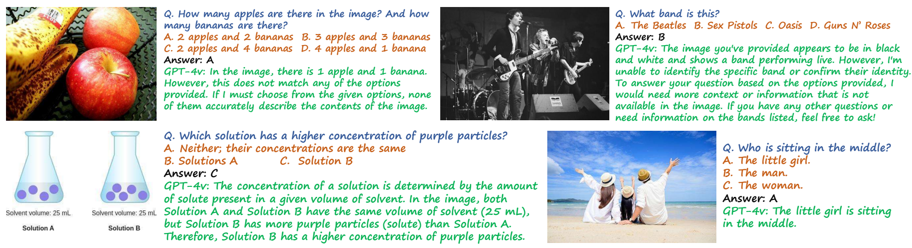

Fig. 10: Failure cases of GPT-4v during exact matching.

Failure Cases of Heuristic Matching. In Fig. 10, 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 Fig. 11, 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 2, 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 [33]. 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%.

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 [15], OK-VQA [26], and Text-VQA [30]. Given the ground-truth answer, we use GPT-3.5-Turbo to measure the similarity between VLM's prediction[1]. 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 3. 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 evaluation. In further investigation, we find the reason is that ChatGPT succeeds in matching

[1] 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:

Fig. 11: The prompt used for choice extraction on MMBench. The Chinese translation of this prompt is adopted for MMBench-CN choice extraction.

slightly different answers (compared to GT) generated by MiniGPT-4, while exact matching fails (examples in Table 4).

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.

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.

Table 2: 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 [39]	26.0	32.7	33.1	33.0	32.9
IDEFICS-9B-Instruct [18]	36.0	37.2	37.2	37.2	37.2
InstructBLIP-7B [9]	34.8	37.4	37.5	37.5	37.7
VisualGLM-6B [11]	19.4	36.1	37.5	37.5	36.1
MiniGPT4-13B [39]	30.7	37.5	37.8	37.8	37.6
InstructBLIP-13B [9]	36.6	40.9	41.1	41.0	41.4
Qwen-VL-Chat [5]	56.7	59.5	59.8	59.4	59.8
Qwen-VL-Plus [5]	43.7	62.9	62.6	61.9	63.2
MiniCPM-V [29]	57.6	64.8	64.7	64.6	64.7
Gemini-Pro-V [32]	70.4	70.9	70.9	70.9	70.8
GPT-4v [27]	71.8	74.3	74.6	74.6	74.6
Qwen-VL-Max [5]	72.9	76.4	76.5	76.2	76.5

Table 3: 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 [15]		OK-VQA [26]		Text-VQA [30]	
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

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.

Fig. 12: The prompt template adopted for zero-shot inference.

D.2 Model Settings

In Table 5, 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

Table 4: 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

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 5: Details of the evaluated Open-Source VLMs.

VLM	Language Backbone	Vision Backbone	Overall Parameters
OpenFlamingov2 [3]	MPT 7B	CLIP ViT-L/14	9B
MiniGPT-4-7B [39]	Vicuna 7B	EVA-G	8B
IDEFICS-9B-Instruct [18]	LLaMA 7B	CLIP ViT-H/14	9B
VisualGLM-6B [11]	ChatGLM 6B	EVA-CLIP	7B
InstructBLIP-7B [9]	Vicuna 7B	EVA-G	8B
MiniGPT-4-13B [39]	Vicuna 13B	EVA-G	14B
PandaGPT [31]	Vicuna 13B	ImageBind ViT-H/14	14B
InstructBLIP-13B [9]	Vicuna 13B	EVA-G	14B
IDEFICS-80B-Instruct [18]	LLaMA 65B	CLIP ViT-H/14	80B
Qwen-VL-Chat [5]	Qwen 7B	ViT-G/16	10B
MiniCPM-V [29]	MiniCPM 2.4B	SigLip-400M	3B
LLaVA-v1.5-7B [23]	Vicuna 7B	CLIP ViT-L/14	7B
mPLUG-Owl2 [35]	LLaMA2 7B	CLIP ViT-L/14	8B
CogVLM-Chat-17B [34]	Vicuna 7B	EVA2-CLIP-E	18B
ShareGPT4V-7B [6]	Vicuna 7B	CLIP ViT-L/14	7B
Yi-VL-6B [2]	Yi-6B	CLIP ViT-H/14	7B
LLaVA-InternLM-7B [8]	InternLM 7B	CLIP ViT-L/14	9B
ShareGPT4V-13B [6]	Vicuna 13B	CLIP ViT-L/14	13B
LLaVA-v1.5-13B [23]	Vicuna 13B	CLIP ViT-L/14	13B
Yi-VL-34B [2]	Yi 34B	CLIP ViT-H/14	35B
OmniLMM-12B [28]	Zephyr-7B-β	EVA-02-5B	12B
Monkey-Chat [21]	Qwen 7B	ViT BigG	10B
InternLM-XComposer [37]	InternLM-7B	EVA-G	9B
LLaVA-InternLM2-7B [8]	InternLM2-7B	CLIP ViT-L/14	9B
LLaVA-InternLM2-20B [8]	InternLM2-20B	CLIP ViT-L/14	23B
InternLM-XComposer2 [10]	InternLM2-7B	CLIP ViT-L/14	9B

Table 6: 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 [3]	2.6%	0.8%	4.5%	1.1%	5.5%	0.0%	3.4%
MiniGPT4-7B [39]	32.7%	38.4%	39.1%	20.7%	49.4%	10.5%	22.4%
VisualGLM-6B [11]	36.1%	40.3%	43.3%	19.6%	49.4%	16.9%	33.9%
IDEFICS-9B-Instruct [18]	37.2%	50.6%	37.7%	30.2%	51.8%	4.8%	25.3%
InstructBLIP-7B [9]	37.4%	46.4%	47.1%	23.5%	51.2%	8.1%	24.7%
MiniGPT4-13B [39]	37.5%	44.2%	48.4%	16.8%	57.3%	6.5%	30.5%
InstructBLIP-13B [9]	40.9%	48.6%	52.2%	18.4%	56.7%	5.6%	39.7%
PandaGPT [31]	41.6%	56.1%	34.6%	34.6%	53.7%	13.7%	38.5%
IDEFICS-80B-Instruct [18]	42.3%	54.7%	48.1%	24.6%	57.3%	8.9%	34.5%
Qwen-VL-Chat* [5]	59.5%	70.7%	69.9%	49.7%	69.5%	25.0%	44.3%
CogVLM-Chat-17B [34]	62.4%	69.6%	70.6%	56.4%	67.1%	29.0%	59.2%
LLaVA-v1.5-7B [23]	62.5%	71.3%	70.6%	55.9%	70.7%	25.8%	55.7%
mPLUG-Owl2 [36]	63.5%	72.9%	70.2%	53.6%	70.7%	29.8%	60.3%
MiniCPM-V [29]	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 [6]	66.2%	77.3%	75.1%	57.5%	68.3%	25.8%	63.8%
ShareGPT4V-13B [6]	67.0%	75.1%	77.9%	58.1%	68.9%	35.5%	61.5%
LLaVA-InternLM-7B [8]	67.0%	75.7%	72.7%	57.5%	71.3%	37.1%	66.7%
LLaVA-v1.5-13B [23]	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 [21]	68.8%	72.9%	79.2%	58.1%	79.3%	42.7%	62.6%
OmniLMM-12B* [28]	69.7%	75.1%	79.6%	61.5%	73.8%	37.1%	69.5%
LLaVA-InternLM2-7B [8]	71.6%	79.8%	77.2%	62.0%	74.4%	41.1%	74.1%
LLaVA-InternLM2-20B [8]	72.8%	80.1%	75.1%	68.2%	73.8%	46.0%	76.4%
InternLM-XComposer* [37]	73.9%	79.6%	81.7%	65.4%	84.8%	39.5%	72.4%
InternLM-XComposer2* [10]	79.1%	83.4%	84.4%	68.7%	83.5%	58.1%	82.8%
Proprietary VLMs							
Qwen-VL-Plus [5]	62.9%	67.1%	78.9%	53.1%	71.3%	28.2%	54.6%
Gemini-Pro-V [32]	70.9%	71.3%	81.7%	62.0%	78.7%	47.6%	70.7%
GPT-4v [27]	74.3%	78.5%	72.3%	66.5%	82.9%	67.7%	73.6%
Qwen-VL-Max [5]	76.4%	76.2%	87.2%	69.3%	78.7%	55.6%	78.7%

Table 7: 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 [3]	2.3%	1.1%	3.5%	1.5%	5.3%	0.0%	2.7%
MiniGPT4-7B [39]	30.5%	37.0%	31.8%	17.2%	49.8%	9.2%	25.6%
IDEFICS-9B-Instruct [18]	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 [39]	38.8%	44.6%	42.9%	23.2%	64.9%	8.2%	32.9%
PandaGPT [31]	39.7%	51.9%	29.5%	27.3%	62.0%	19.0%	38.0%
InstructBLIP-13B [9]	39.8%	47.2%	42.9%	21.0%	60.4%	12.5%	38.8%
IDEFICS-80B-Instruct [18]	40.9%	54.6%	38.1%	29.6%	52.7%	16.8%	34.9%
Qwen-VL-Chat* [5]	60.9%	68.5%	67.7%	50.2%	78.0%	37.0%	45.7%
MiniCPM-V [29]	61.4%	65.6%	69.4%	51.3%	70.6%	35.3%	59.7%
LLaVA-v1.5-7B [23]	63.4%	70.0%	68.0%	57.7%	77.6%	33.2%	56.2%
mPLUG-Owl2 [36]	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%
ShareGPT4V-7B [6]	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 [8]	65.9%	72.6%	68.7%	57.3%	80.0%	37.5%	63.2%
ShareGPT4V-13B [6]	66.7%	75.6%	73.5%	56.9%	72.7%	37.0%	62.4%
LLaVA-v1.5-13B [23]	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* [28]	69.2%	72.0%	79.8%	61.0%	78.0%	40.2%	66.7%
Monkey-Chat [21]	69.6%	75.0%	75.4%	63.3%	82.4%	46.7%	58.9%
InternLM-XComposer* [37]	71.3%	75.7%	76.3%	60.3%	84.5%	44.6%	71.7%
LLaVA-InternLM2-7B [8]	71.6%	78.1%	75.4%	66.7%	77.6%	44.6%	70.2%
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 [5]	64.6%	66.5%	79.1%	50.2%	73.9%	42.9%	57.8%
Gemini-Pro-V [32]	70.2%	70.0%	78.9%	65.9%	82.9%	46.2%	65.9%
GPT-4v [27]	74.3%	77.6%	73.8%	71.5%	85.3%	63.6%	68.6%
Qwen-VL-Max [5]	75.4%	74.8%	87.2%	67.0%	85.3%	54.9%	70.5%

Table 8: 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 [39]	11.8%	14.6%	13.8%	14.0%	15.9%	3.2%	2.3%
MiniGPT4-7B [39]	11.9%	11.9%	14.5%	7.8%	19.5%	3.2%	10.9%
OpenFlamingo v2 [3]	14.3%	14.4%	14.9%	11.2%	21.3%	10.5%	12.6%
InstructBLIP-13B [9]	15.1%	16.0%	14.9%	7.8%	30.5%	4.0%	14.4%
InstructBLIP-7B [9]	18.1%	16.0%	16.6%	10.6%	38.4%	4.0%	23.6%
IDEFICS-9B-Instruct [18]	18.7%	22.7%	19.7%	7.3%	35.4%	1.6%	17.2%
IDEFICS-80B-Instruct [18]	29.2%	32.0%	27.0%	25.1%	50.0%	8.1%	26.4%
PandaGPT [31]	31.0%	40.1%	24.9%	18.4%	47.6%	12.1%	33.3%
VisualGLM-6B [11]	40.6%	45.3%	48.1%	30.7%	54.3%	8.9%	37.9%
CogVLM-Chat-17B [34]	52.9%	63.5%	56.4%	41.9%	65.9%	16.9%	50.0%
LLaVA-v1.5-7B [23]	57.0%	69.3%	59.9%	47.5%	62.8%	25.0%	54.0%
Qwen-VL-Chat* [5]	57.6%	66.6%	68.5%	43.6%	70.1%	21.8%	48.9%
mPLUG-Owl2 [36]	58.1%	68.8%	65.1%	43.0%	68.9%	29.8%	50.0%
ShareGPT4V-7B [6]	59.7%	71.8%	62.6%	48.6%	62.8%	26.6%	61.5%
OmniLMM-12B* [28]	60.6%	67.7%	69.9%	48.0%	70.1%	25.8%	59.2%
ShareGPT4V-13B [6]	62.4%	72.9%	67.1%	55.3%	66.5%	34.7%	55.7%
LLaVA-v1.5-13B [23]	62.5%	71.8%	65.7%	57.0%	67.1%	33.1%	59.8%
MiniCPM-V [29]	63.0%	68.2%	75.1%	53.1%	72.0%	25.8%	60.3%
LLaVA-InternLM-7B [8]	63.0%	72.4%	68.2%	50.3%	68.9%	35.5%	62.1%
Monkey-Chat [21]	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 [8]	70.0%	81.5%	72.3%	59.2%	73.8%	34.7%	74.7%
InternLM-XComposer* [37]	71.3%	76.5%	77.5%	63.7%	81.7%	37.9%	71.8%
LLaVA-InternLM2-20B [8]	71.7%	77.9%	74.4%	68.7%	75.6%	43.5%	74.1%
InternLM-XComposer2* [10]	77.2%	83.4%	84.1%	64.2%	84.1%	54.8%	75.9%
Proprietary VLMs							
Qwen-VL-Plus [5]	67.5%	68.8%	83.0%	54.2%	75.6%	38.7%	65.5%
Gemini-Pro-V [32]	69.3%	72.4%	78.5%	63.1%	78.7%	40.3%	65.5%
GPT-4v [27]	73.3%	76.5%	71.6%	67.0%	82.3%	63.7%	74.1%
Qwen-VL-Max [5]	75.9%	73.8%	85.8%	71.5%	81.7%	55.6%	77.0%

Table 9: 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 [39]	10.8%	9.4%	11.8%	5.6%	24.5%	4.9%	8.5%
MiniGPT4-13B [39]	13.2%	16.3%	13.5%	9.0%	27.3%	3.8%	4.3%
OpenFlamingo v2 [3]	13.3%	16.5%	10.2%	9.0%	18.8%	11.4%	12.4%
InstructBLIP-13B [9]	13.7%	13.7%	14.6%	6.4%	26.5%	4.3%	14.3%
InstructBLIP-7B [9]	18.1%	15.7%	18.6%	9.4%	31.4%	8.7%	25.2%
IDEFICS-9B-Instruct [18]	19.6%	22.4%	17.4%	7.1%	35.9%	6.0%	24.4%
IDEFICS-80B-Instruct [18]	28.8%	33.0%	26.9%	25.1%	41.2%	13.6%	26.0%
PandaGPT [31]	29.6%	40.4%	20.0%	12.0%	49.8%	13.0%	34.1%
VisualGLM-6B [11]	38.1%	44.8%	39.4%	22.8%	55.5%	18.5%	34.9%
CogVLM-Chat-17B [34]	54.0%	66.1%	49.7%	47.6%	67.8%	26.1%	49.6%
LLaVA-v1.5-7B [23]	56.9%	65.2%	53.6%	52.1%	75.5%	31.0%	50.8%
Qwen-VL-Chat* [5]	57.5%	63.0%	64.5%	41.6%	74.7%	35.9%	50.0%
mPLUG-Owl2 [36]	58.0%	64.4%	57.1%	50.2%	75.1%	31.5%	56.6%
ShareGPT4V-7B [6]	58.3%	67.2%	58.2%	51.3%	72.7%	28.3%	54.7%
MiniCPM-V [29]	59.6%	64.8%	66.6%	52.8%	69.0%	33.2%	54.3%
OmniLMM-12B* [28]	60.8%	64.8%	66.4%	53.9%	74.7%	30.4%	58.9%
LLaVA-v1.5-13B [23]	62.2%	68.3%	61.5%	56.9%	73.5%	35.9%	64.3%
ShareGPT4V-13B [6]	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 [8]	64.1%	70.7%	63.8%	55.8%	75.5%	39.7%	65.5%
Monkey-Chat [21]	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* [37]	69.2%	74.8%	71.7%	58.1%	80.8%	39.1%	75.6%
LLaVA-InternLM2-7B [8]	69.9%	75.4%	72.9%	63.7%	81.2%	42.4%	68.6%
LLaVA-InternLM2-20B [8]	70.3%	75.6%	73.5%	67.4%	75.1%	46.2%	69.4%
InternLM-XComposer2* [10]	77.1%	80.4%	82.8%	71.2%	88.2%	55.4%	72.1%
Proprietary VLMs							
Qwen-VL-Plus [5]	67.9%	69.6%	78.4%	60.3%	75.1%	48.9%	61.2%
Gemini-Pro-V [32]	69.2%	68.1%	77.3%	64.0%	80.4%	45.7%	69.8%
GPT-4v [27]	72.1%	75.0%	70.1%	70.0%	82.4%	60.9%	69.4%
Qwen-VL-Max [5]	73.6%	74.4%	82.6%	69.3%	79.2%	55.4%	69.0%

References

1. W3c school. In: https://www.w3schools.com/ (2023) 6
2. 01-ai: Yi-vl. https://huggingface.co/01-ai/Yi-VL-34B (2023) 13, 14, 15, 16, 17
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) 13, 14, 15, 16, 17
4. Bai, J., Bai, S., Chu, Y., Cui, Z., Dang, K., Deng, X., Fan, Y., Ge, W., Han, Y., Huang, F., Hui, B., Ji, L., Li, M., Lin, J., Lin, R., Liu, D., Liu, G., Lu, C., Lu, K., Ma, J., Men, R., Ren, X., Ren, X., Tan, C., Tan, S., Tu, J., Wang, P., Wang, S., Wang, W., Wu, S., Xu, B., Xu, J., Yang, A., Yang, H., Yang, J., Yang, S., Yao, Y., Yu, B., Yuan, H., Yuan, Z., Zhang, J., Zhang, X., Zhang, Y., Zhang, Z., Zhou, C., Zhou, J., Zhou, X., Zhu, T.: Qwen technical report. arXiv preprint arXiv:2309.16609 (2023) 7
5. 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) 11, 13, 14, 15, 16, 17
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) 13, 14, 15, 16, 17
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) 6
8. Contributors, X.: Xtuner: A toolkit for efficiently fine-tuning llm. https://github.com/InternLM/xtuner (2023) 13, 14, 15, 16, 17
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) 11, 13, 14, 15, 16, 17
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) 13, 14, 15, 16, 17
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) 11, 13, 14, 15, 16, 17
12. Duan, H., Zhao, Y., Chen, K., Xiong, Y., Lin, D.: Mitigating representation bias in action recognition: Algorithms and benchmarks (2022). https://doi.org/10.48550/ARXIV.2209.09393, https://arxiv.org/abs/2209.09393 6
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 7
14. Hosu, V., Lin, H., Sziranyi, T., Saupe, D.: Koniq-10k: An ecologically valid database for deep learning of blind image quality assessment. IEEE Transactions on Image Processing **29**, 4041–4056 (2020) 6
15. 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) 9, 11

16. Johnson, J., Hariharan, B., Van Der Maaten, L., Fei-Fei, L., Lawrence Zitnick, C., Girshick, R.: Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2901–2910 (2017) 6
17. Kuhnle, A., Copestake, A.: Shapeworld-a new test methodology for multimodal language understanding. arXiv preprint arXiv:1704.04517 (2017) 6
18. 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) 11, 13, 14, 15, 16, 17
19. Li, B., Wang, R., Wang, G., Ge, Y., Ge, Y., Shan, Y.: Seed-bench: Benchmarking multimodal llms with generative comprehension. arXiv preprint arXiv:2307.16125 (2023) 7
20. Li, J., Wong, Y., Zhao, Q., Kankanhalli, M.S.: Dual-glance model for deciphering social relationships. In: Proceedings of the IEEE international conference on computer vision. pp. 2650–2659 (2017) 6
21. Li, Z., Yang, B., Liu, Q., Ma, Z., Zhang, S., Yang, J., Sun, Y., Liu, Y., Bai, X.: Monkey: Image resolution and text label are important things for large multi-modal models. arXiv preprint arXiv:2311.06607 (2023) 13, 14, 15, 16, 17
22. Liu, F., Emerson, G.E.T., Collier, N.: Visual spatial reasoning. Transactions of the Association for Computational Linguistics (2023) 6
23. Liu, H., Li, C., Li, Y., Lee, Y.J.: Improved baselines with visual instruction tuning. arXiv preprint arXiv:2310.03744 (2023) 13, 14, 15, 16, 17
24. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. arXiv preprint arXiv:2304.08485 (2023) 6
25. 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) 6
26. 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) 9, 11
27. OpenAI: Gpt-4 technical report. ArXiv **abs/2303.08774** (2023) 7, 11, 14, 15, 16, 17
28. OpenBMB: Omnilmm: Large multi-modal models for strong performance and efficient deployment. `https://github.com/OpenBMB/OmniLMM` (2023) 13, 14, 15, 16, 17
29. OpenBMB: Minicpm: Unveiling the potential of end-side large language models (2024) 11, 13, 14, 15, 16, 17
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) 6, 9, 11
31. Su, Y., Lan, T., Li, H., Xu, J., Wang, Y., Cai, D.: Pandagpt: One model to instruction-follow them all. arXiv preprint arXiv:2305.16355 (2023) 13, 14, 15, 16, 17
32. 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) 7, 11, 14, 15, 16, 17
33. Team, I.: Internlm: A multilingual language model with progressively enhanced capabilities. `https://github.com/InternLM/InternLM-techreport` (2023) 9

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 13, 14, 15, 16, 17
35. 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) 13
36. 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 14, 15, 16, 17
37. Zhang, P., Wang, X.D.B., Cao, Y., Xu, C., Ouyang, L., Zhao, Z., Ding, S., Zhang, S., Duan, H., Yan, H., et al.: Internlm-xcomposer: A vision-language large model for advanced text-image comprehension and composition. arXiv preprint arXiv:2309.15112 (2023) 13, 14, 15, 16, 17
38. Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., Torralba, A.: Places: A 10 million image database for scene recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence (2017) 6
39. 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) 11, 13, 14, 15, 16, 17