MME: A Comprehensive Evaluation Benchmark for Multimodal Large Language Models

Chaoyou Fu[1], **Peixian Chen**[1], **Yunhang Shen**[1], **Yulei Qin**[1], **Mengdan Zhang**[1], **Xu Lin**[1]
Zhenyu Qiu[2], **Wei Lin**[2], **Jinrui Yang**[1], **Xiawu Zheng**[2], **Ke Li**[1,*], **Xing Sun**[1], **Rongrong Ji**[2]
[1]Tencent Youtu Lab [2]Xiamen University
bradyfu24@gmail.com, {tristanli, winfredsun}@tencent.com

Abstract

Multimodal Large Language Model (MLLM) relies on the powerful LLM to perform multimodal tasks, showing amazing emergent abilities in recent studies, such as writing poems based on an image. However, it is difficult for these case studies to fully reflect the performance of MLLM, lacking a comprehensive evaluation. In this paper, we fill in this blank, presenting the first **MLLM E**valuation benchmark **MME**[2]. It measures both perception and cognition abilities on a total of 14 subtasks. In order to avoid data leakage that may arise from direct use of public datasets for evaluation, the annotations of instruction-answer pairs are all manually designed. The concise instruction design allows us to fairly compare MLLMs, instead of struggling in prompt engineering. Besides, with such an instruction, we can also easily carry out quantitative statistics. A total of **12** advanced MLLMs are comprehensively evaluated on our MME, which not only suggests that existing MLLMs still have a large room for improvement, but also reveals the potential directions for the subsequent model optimization.

1 Introduction

The thriving of Large Language Model (LLM) has paved a new road to the multimodal field, i.e., Multimodal Large Language Model (MLLM) [40, 13, 26, 30, 19]. It refers to using LLM as a brain to process multimodal information and give reasoning results. Equipped with the powerful LLM, MLLM is expected to address more complex multi-modal tasks [19, 47, 41]. The three representative abilities of LLM [52], including In-Context Learning (ICL) [16], instruction following [43], and Chain-of-Thought (CoT) [45], are also manifested in multimodality. For example, Flamingo [13] turns on multimodal ICL, which can adapt to new tasks by giving a few examples. MiniGPT-4 [56] implements GPT-4[40]-like instruction following capabilities, such as converting images into corresponding website codes, by introducing multimodal instruction tuning. PaLM-E [19] achieves amazing OCR-free math reasoning via CoT. These emergent abilities of MLLM are exciting and imply that a new dawn has broken in artificial intelligence.

Although these models exhibit surprising conversational capabilities when conducting everyday chats, we still know little about how well they quantitatively perform in various aspects. The existing three common quantitative evaluation manners for MLLMs have their limitations that are difficult to comprehensively evaluate performance. Specifically, the first manner [49, 18, 44] evaluates on existing traditional multimodal datasets, such as image caption [17] and VQA [23, 39, 35]. However, on the one hand, it may be hard to reflect the emergent abilities of MLLM on these datasets. On the

[*]Ke Li is the corresponding author.

[2]MME is collected by Xiamen University for academic research only. This is the v1 version of MME, which will be updated as MLLMs evolve. The data application manner and online leaderboards are released at `https://github.com/BradyFU/Awesome-Multimodal-Large-Language-Models/tree/Evaluation`.

Preprint. Under review.

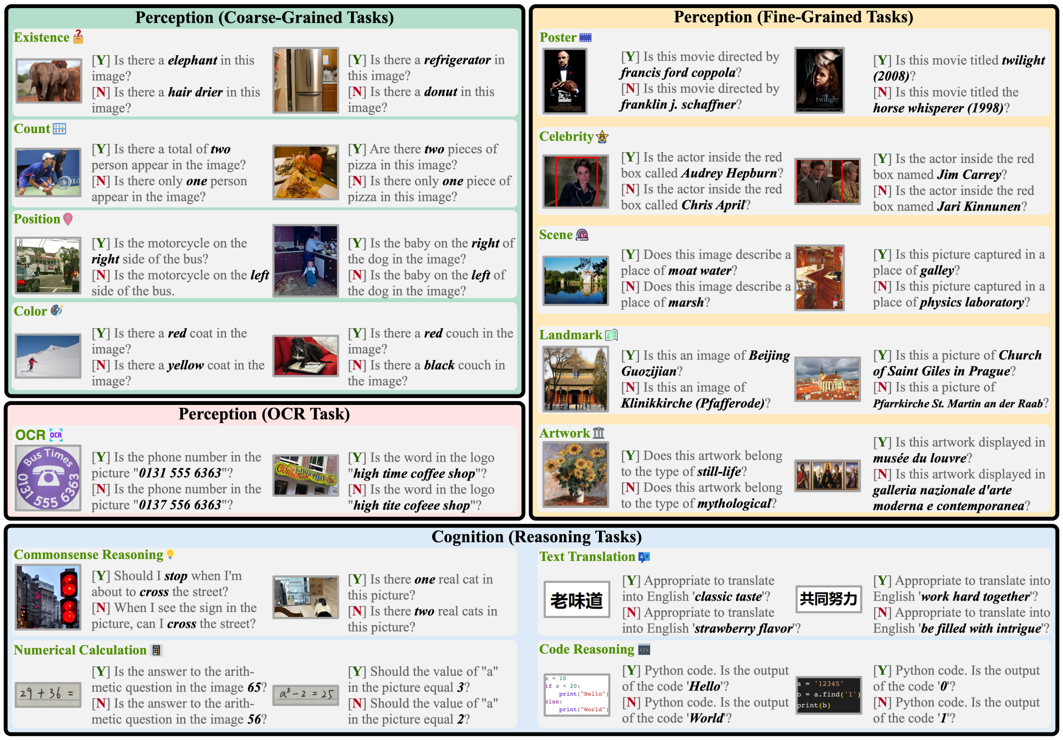

Figure 1: Diagram of our MME benchmark. It evaluates MLLM from both perception and cognition, including a total of 14 subtasks. Each image corresponds to two questions whose answers are marked yes [Y] and no [N], respectively. The instruction consists of a question followed by "Please answer yes or no". It is worth noting that all instructions are manually designed.

other hand, since the training sets of large models are no longer unified, it is difficult to guarantee that all MLLMs have not used the testing set for training. The second manner [50] is to collect data for an open-ended evaluation, but either the data is unavailable to public by now [54] or the amount is small (only 50 images) [50]. The third manner focuses on one aspect of MLLMs, such as object hallucination [31] or adversarial robustness [53], which is powerless to comprehensive evaluation.

In light of these concerns, a new comprehensive evaluation benchmark is urgently needed to match the flourish of MLLMs. We argue that a universal comprehensive evaluation benchmark should have the following four characteristics: (1) It should cover as much as possible, including both perception and cognition abilities. The former refers to recognizing the specific object, such as its existence, count, position, and color. The latter refers to compositing the perception information and the knowledge in LLM to deduce more complex answers. It is obvious that the former is the premise of the latter. (2) Its data or annotations should not come from existing publicly available datasets as much as possible, avoiding the risk of data leakage. (3) Its instructions should be as concise as possible and in line with human cognition. Although instruction design may have a large impact on the output, all models should be tested under the same unified instructions for fair comparison. A good MLLM should be able to generalize to such concise instructions. (4) The responses of MLLM to the instructions should be intuitive and convenient for quantitative analysis. The open-ended answer of MLLM poses significant challenges to the quantization. Existing methods tend to use GPT or manual scoring [28, 33, 50], but there may be problems of inaccuracy and subjectivity.

To this end, we collect a comprehensive MLLM Evaluation benchmark, named as MME, which meets the above four characteristics at the same time:

- MME covers the examination of perception and cognition abilities. Apart from OCR, the perception includes the recognition of coarse-grained and fine-grained objects. The former identifies the existence, count, position, and color of objects. The latter recognizes movie posters, celebrities, scenes, landmarks, and artworks. The cognition includes commonsense reasoning, numerical calculation, text translation, and code reasoning. The total number of subtasks is up to 14, as shown in Fig. 1.

- All instruction-answer pairs are manually constructed. For the few public datasets involved in our study, we only use images without directly relying on their original annotations. Meanwhile, we make efforts to collect data through real photographs and image generation.
- The instructions of MME are designed concisely to avoid the impact of prompt engineering on the model output. We argue that a good MLLM should be able to generalize to such simple and frequently used instructions, which are fair to all models. Please see Fig. 1 for the specific instruction of each subtask.
- Benefitting from our instruction design "please answer yes or no", we can easily perform quantitative statistics based on the "yes" or "no" output of MLLMs, which is accurate and objective. It should be noted that we have also tried to design instructions with multiple choice questions, but find that it may beyond the capabilities of current MLLMs to follow complex instructions.

We conduct massive experiments to evaluate the zero-shot performance of 12 advanced MLLMs on 14 subtasks. The evaluated MLLMs include BLIP-2 [30], LLaVA [33], MiniGPT-4 [56], mPLUG-Owl [50], LLaMA-Adapter-v2 [20], Otter [29], Multimodal-GPT [22], InstructBLIP [18], VisualGLM-6B [12], PandaGPT [42], ImageBind-LLM [2], and LaVIN [36]. As displayed in Fig. 2 that consists of 2 overall leaderboards (perception and cognition) and 14 individual leaderboards, these MLLMs show clear discrepancies in our MME evaluation benchmark. Fig. 3 also provides a comparison from the other perspective. More importantly, we have summarized four prominent problems exposed in experiments, including inability to follow basic instructions, a lack of basic perception and reasoning, as well as object hallucination, as shown in Fig. 4. These findings are instructive for the subsequent model optimization.

In summary, the contributions of this work are as follows: (1) We propose a new benchmark MME to meet the urgent need of MLLM evaluation. As far as we know, MME is the first comprehensive MLLM evaluation benchmark[3]. (2) We conduct abundant experiments on a total of 14 subtasks to comprehensively evaluate 12 up-to-date MLLMs. (3) We summarize the exposed problems in experiments, proving guidance for the evolution of MLLM.

2 MME Evaluation Suite

2.1 Instruction Design

In order to facilitate quantitative performance statistics, the orientation of our instruction design is to let the model to answer "yes" or "no". As a result, the instruction consists of two parts, including a concise question and a description "Please answer yes or no." For each test image, we manually design two instructions, where the discrepancy lies in the questions. The ground truth answer of the first question is "yes" and that of the second question is "no", as shown in Fig. 1. When MLLM answers both of these questions correctly, it appears more confident that the MLLM actually comprehends the image and the corresponding knowledge behind it.

2.2 Evaluation Metric

Since the output of the model is limited to two types ("yes" or "no"), it is convenient to measure the metrics of accuracy and accuracy+. The former is calculated based on each question, while the latter is based on each image where both of the two questions need to be answered correctly. The random accuracies of the two metrics are equal to 50% and 25%, respectively. It can be seen that accuracy+ is a stricter measurement but also better reflects the comprehensive understanding degree of the model to the image. In addition, we calculate the score of a subtask based on the sum of accuracy and accuracy+. The perception score is the sum of scores of all perception subtasks. The cognition score is calculated in the same way.

[3]The publication date of our first version is June 23, 2023. The concurrent works are LAMM-Benchmark [51] and LVLM-eHub [48], which are rather different from our MME. We recommend reading the two works as well to enrich the understanding of MLLM evaluation.

Rank	Model	Score
🥇	BLIP-2	1293.84
🥈	InstructBLIP	1212.82
🥉	LLaMA-Adapter V2	972.67
4	mPLUG-Owl	967.35
5	LaVIN	963.61
6	MiniGPT-4	866.58
7	ImageBind_LLM	775.77
8	VisualGLM-6B	705.31
9	Multimodal-GPT	654.73
10	PandaGPT	642.59
11	LLaVA	502.82
12	Otter	483.73

(1) **Perception**

Rank	Model	Score
🥇	MiniGPT-4	292.14
🥈	InstructBLIP	291.79
🥉	BLIP-2	290.00
4	mPLUG-Owl	276.07
5	LaVIN	249.64
6	LLaMA-Adapter V2	248.93
7	PandaGPT	228.57
8	Multimodal-GPT	226.79
9	LLaVA	214.64
10	ImageBind_LLM	213.57
11	VisualGLM-6B	181.79
12	Otter	136.07

(2) **Cognition**

Rank	Model	Score
🥇	InstructBLIP	185.00
🥇	LaVIN	185.00
🥈	BLIP-2	160.00
🥉	ImageBind_LLM	128.33
4	mPLUG-Owl	120.00
4	LLaMA-Adapter V2	120.00
5	MiniGPT-4	115.00
6	VisualGLM-6B	85.00
7	PandaGPT	70.00
8	Multimodal-GPT	61.67
9	LLaVA	50.00
10	Otter	48.33

(3) Existence

Rank	Model	Score
🥇	InstructBLIP	143.33
🥈	BLIP-2	135.00
🥉	MiniGPT-4	123.33
4	LaVIN	88.33
5	ImageBind_LLM	60.00
6	Multimodal-GPT	55.00
7	mPLUG-Owl	50.00
7	LLaMA-Adapter V2	50.00
7	VisualGLM-6B	50.00
7	Otter	50.00
7	PandaGPT	50.00
7	LLaVA	50.00

(4) Count

Rank	Model	Score
🥇	MiniGPT-4	81.67
🥈	BLIP-2	73.33
🥉	InstructBLIP	66.67
4	LaVIN	63.33
5	Multimodal-GPT	58.33
6	mPLUG-Owl	50.00
6	Otter	50.00
6	PandaGPT	50.00
6	LLaVA	50.00
7	LLaMA-Adapter V2	48.33
7	VisualGLM-6B	48.33
8	ImageBind_LLM	46.67

(5) Position

Rank	Model	Score
🥇	InstructBLIP	153.33
🥈	BLIP-2	148.33
🥉	MiniGPT-4	110.00
4	LLaMA-Adapter V2	75.00
4	LaVIN	75.00
5	ImageBind_LLM	73.33
6	Multimodal-GPT	68.33
7	mPLUG-Owl	55.00
7	VisualGLM-6B	55.00
7	Otter	55.00
7	LLaVA	55.00
8	PandaGPT	50.00

(6) Color

Rank	Model	Score
🥇	BLIP-2	141.84
🥈	mPLUG-Owl	136.05
🥉	InstructBLIP	123.81
4	LLaMA-Adapter V2	99.66
5	LaVIN	79.59
6	PandaGPT	76.53
7	VisualGLM-6B	65.99
8	ImageBind_LLM	64.97
9	Multimodal-GPT	57.82
10	MiniGPT-4	55.78
11	LLaVA	50.00
12	Otter	44.90

(7) Poster

Rank	Model	Score
🥇	BLIP-2	105.59
🥈	InstructBLIP	101.18
🥉	mPLUG-Owl	100.29
4	LLaMA-Adapter V2	86.18
5	ImageBind_LLM	76.47
6	Multimodal-GPT	73.82
7	MiniGPT-4	65.29
8	PandaGPT	57.06
9	VisualGLM-6B	53.24
10	Otter	50.00
11	LLaVA	48.82
12	LaVIN	47.35

(8) Celebrity

Rank	Model	Score
🥇	InstructBLIP	153.00
🥈	LLaMA-Adapter V2	148.50
🥉	VisualGLM-6B	146.25
4	BLIP-2	145.25
5	LaVIN	136.75
6	mPLUG-Owl	135.50
7	PandaGPT	118.00
8	ImageBind_LLM	113.25
9	MiniGPT-4	95.75
10	Multimodal-GPT	68.00
11	LLaVA	50.00
12	Otter	44.25

(9) Scene

Rank	Model	Score
🥇	mPLUG-Owl	159.25
🥈	LLaMA-Adapter V2	150.25
🥉	BLIP-2	138.00
4	LaVIN	93.50
5	VisualGLM-6B	83.75
6	InstructBLIP	79.75
7	Multimodal-GPT	69.75
7	PandaGPT	69.75
8	MiniGPT-4	69.00
9	ImageBind_LLM	62.00
10	LLaVA	50.00
11	Otter	49.50

(10) Landmark

Rank	Model	Score
🥇	BLIP-2	136.50
🥈	InstructBLIP	134.25
🥉	mPLUG-Owl	96.25
4	LaVIN	87.25
5	VisualGLM-6B	75.25
6	ImageBind_LLM	70.75
7	LLaMA-Adapter V2	69.75
8	Multimodal-GPT	59.50
9	MiniGPT-4	55.75
10	PandaGPT	51.25
11	LLaVA	49.00
12	Otter	41.75

(11) Artwork

Rank	Model	Score
🥇	LLaMA-Adapter V2	125.00
🥈	BLIP-2	110.00
🥉	LaVIN	107.50
4	MiniGPT-4	95.00
5	Multimodal-GPT	82.50
6	ImageBind_LLM	80.00
7	InstructBLIP	72.50
8	mPLUG-Owl	65.00
9	Otter	50.00
10	PandaGPT	50.00
11	LLaVA	50.00
12	VisualGLM-6B	42.50

(l2) OCR

Rank	Model	Score
🥇	InstructBLIP	129.29
🥈	BLIP-2	110.00
🥉	LaVIN	87.14
4	LLaMA-Adapter V2	81.43
5	mPLUG-Owl	78.57
6	PandaGPT	73.57
7	MiniGPT-4	72.14
8	LLaVA	57.14
9	Multimodal-GPT	49.29
10	ImageBind_LLM	48.57
11	VisualGLM-6B	39.29
12	Otter	38.57

(13) Commonsense Reasoning

Rank	Model	Score
🥇	LaVIN	65.00
🥈	LLaMA-Adapter V2	62.50
🥈	Multimodal-GPT	62.50
🥉	mPLUG-Owl	60.00
4	MiniGPT-4	55.00
4	ImageBind_LLM	55.00
5	PandaGPT	50.00
5	LLaVA	50.00
6	VisualGLM-6B	45.00
7	BLIP-2	40.00
7	InstructBLIP	40.00
8	Otter	20.00

(14) Numerical Calculation

Rank	Model	Score
🥇	mPLUG-Owl	80.00
🥈	BLIP-2	65.00
🥈	InstructBLIP	65.00
🥉	Multimodal-GPT	60.00
4	PandaGPT	57.50
5	LLaVA	57.50
6	MiniGPT-4	55.00
7	ImageBind_LLM	50.00
7	LLaMA-Adapter V2	50.00
7	VisualGLM-6B	50.00
8	LaVIN	47.50
9	Otter	27.50

(15) Text Translation

Rank	Model	Score
🥇	MiniGPT-4	110.00
🥈	BLIP-2	75.00
🥉	ImageBind_LLM	60.00
4	mPLUG-Owl	57.50
4	InstructBLIP	57.50
5	LLaMA-Adapter V2	55.00
5	Multimodal-GPT	55.00
6	Otter	50.00
6	LLaVA	50.00
6	LaVIN	50.00
7	VisualGLM-6B	47.50
7	PandaGPT	47.50

(16) Code Reasoning

Figure 2: Leaderboards on our MME benchmark. (1) and (2) are the overall leaderboards of perception and cognition respectively, in which the full score of the former is 2000 and that of the latter is 800. (3)-(16) are the leaderboards of the 14 subtasks with the full score of 200. The score is the sum of the accuracy and the accuracy+ in Tables. 1, 2, and 3.

2.3 Data Collection

2.3.1 Perception Tasks

We argue that perception is one of the most fundamental capabilities of MLLM, and the lack of perception will easily lead to the object hallucination problem [31]. That is, MLLM will answer

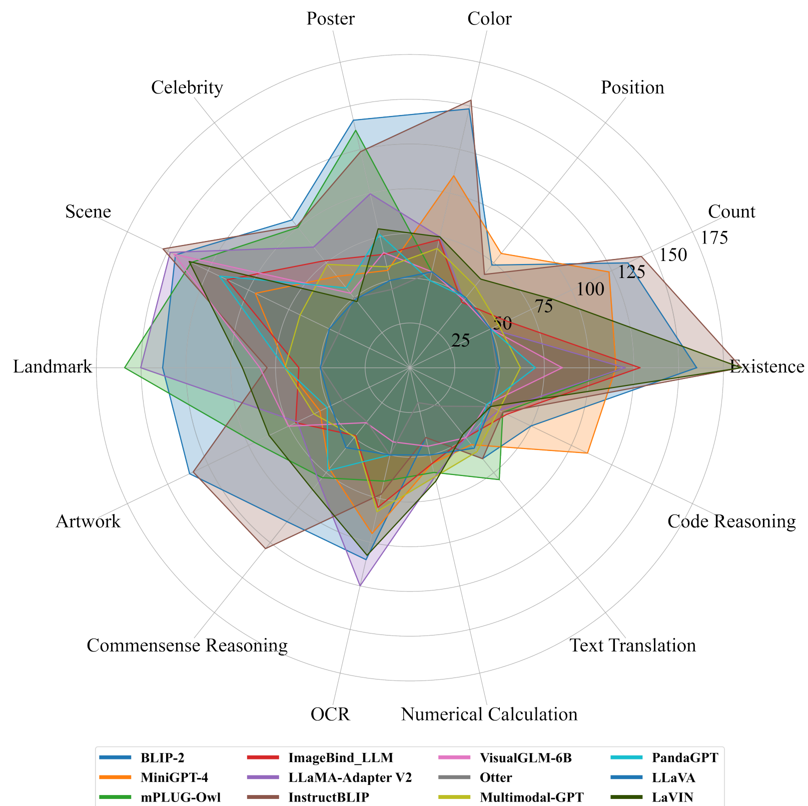

Figure 3: Comparison of 12 advanced MLLMs on 14 subtasks. The full score of each subtask is 200.

questions based on its own fantasies rather than based on the realistic content of the image, as displayed in Fig. 4.

Coarse-Grained Recognition. The contents of coarse-grained recognition include the existence of common objects, and their count, color, and position. The images are sampled from COCO [32], but the instruction-answer pairs are all manually constructed, rather than directly using publicly available annotations. Even if MLLMs have seen these COCO images, our manually prepared pairs are not presented in their training sets. This requires MLLM to be able to understand the instructions and infer corresponding answers. In each perception subtask of existence, count, color, and position, we prepare 30 images with 60 instruction-answer pairs.

Fine-Grained Recognition. The fine-grained recognition is more about testing the knowledge resources of MLLM. The subtasks consist of recognizing movie posters, celebrities, scenes, landmarks, and artworks, containing 147, 170, 200, 200, and 200 images respectively. For the celebrities, we plot a red box to a person with a clearly visible face in the image, and the corresponding instruction is “Is the actor inside the red box named [celebrity name]? Please answer yes or no.” Similar with the above coarse-grained recognition, the images of these subtasks are from publicly available datasets [25, 37, 38, 55, 46] and all of the instructions are manually designed.

OCR. Optical Character Recognition (OCR) is also a foundational capability of MLLM, serving for subsequent text-based tasks such as text translation and text understanding. The images are from [34] and the instruction-answer pairs are manually designed. Considering that MLLM is still in its infancy, we only choose the relatively simple samples in this version of MME. The numbers of image and instruction-answer pairs are 20 and 40, respectively.

2.3.2 Cognition Tasks

We evaluate if any MLLM can carry out further logical reasoning after perceiving the image, which is the most fascinating aspect of MLLM over previous traditional methods. In order to infer the

Model	Existence		Count		Position		Color		OCR	
	ACC	ACC+	ACC	ACC+	ACC	ACC+	ACC	ACC+	ACC	ACC+
BLIP-2	86.67	73.33	75.00	60.00	56.67	16.67	81.67	66.67	70.00	40.00
MiniGPT-4	65.00	50.00	66.67	56.67	48.33	33.33	66.67	43.33	60.00	35.00
mPLUG-Owl	73.33	46.67	50.00	0.00	50.00	0.00	51.67	3.33	55.00	10.00
ImageBind-LLM	75.00	53.33	50.00	10.00	43.33	3.33	56.67	16.67	60.00	20.00
LLaMA-AdapterV2	73.33	46.67	50.00	0.00	48.33	0.00	58.33	16.67	75.00	50.00
InstructBLIP	95.00	90.00	80.00	63.33	53.33	13.33	83.33	70.00	57.50	15.00
VisualGLM-6B	61.67	23.33	50.00	0.00	48.33	0.00	51.67	3.33	42.50	0.00
Otter	48.33	0.00	50.00	0.00	50.00	0.00	51.67	3.33	50.00	0.00
Multimodal-GPT	48.33	13.33	48.33	6.67	45.00	13.33	55.00	13.33	57.50	25.00
PandaGPT	56.67	13.33	50.00	0.00	50.00	0.00	50.00	0.00	50.00	0.00
LLaVA	50.00	0.00	50.00	0.00	50.00	0.00	51.67	3.33	50.00	0.00
LaVIN	95.00	90.00	61.67	26.67	53.33	10.00	58.33	16.67	67.50	40.00

Table 1: Evaluation results (%) of coarse-grained recognition and OCR.

correct answer, MLLM needs to follow the instruction, perceive the contents of the image, and invoke the knowledge reserved in LLM, which is much more challenging than the single perception tasks. Examples of the following subtasks are shown in Fig. 1.

Commonsense Reasoning. Unlike the ScienceQA dataset [35] that requires specialized knowledge, the commonsense refers to the basic knowledge in daily life. For example, given a photo of a down jacket, asking MLLM whether it is appropriate to wear the cloth when it is cold (or hot). These are basic knowledge that humans can judge instantly without complex step-by-step reasoning. Therefore, we expect MLLM to perform well in a zero-short setting. The images are all manually photographed or generated by diffusion models, and the instruction-answer pairs are all manually designed. There are a total of 70 images and 140 instruction-answer pairs.

Numerical Calculation. It requires MLLM to be able to read the arithmetic problem in the image and output the answer in an end to end way, which has been demonstrated in [26]. In this version, we only consider relatively easy arithmetic problems, such as addition, subtraction, multiplication, and division. There are a total of 20 images and 40 instruction-answer pairs. The images are all manually taken, and the instruction-answer pairs are all manually designed.

Text Translation. Considering that the MLLM [12] supports both English and Chinese, we set the text translation task. It requires MLLM to translate the Chinese written in an image to the corresponding English. In this version, we only design basic translation problems, which will be updated according to the development of MLLM in the future. The images of this part are all manually taken, and the instruction-answer pairs are all manually designed. There are a total of 20 images and 40 instruction-answer pairs.

Code Reasoning. It requires MLLM to read the code in the images and automatically complete logical operation inside the code. A similar task that writes website code based on an image has been demonstrated in [56]. The images are all manually taken, and the instruction-answer pairs are all manually designed. We only set basic code problems in this version. There are in total 20 images and 40 instruction-answer pairs.

3 Experiments

In this section, we conduct massive experiments on our MME benchmark to evaluate a total of 12 open-source MLLMs, including BLIP-2, LLaVA, MiniGPT-4, mPLUG-Owl, LLaMA-Adapter-v2, Otter, Multimodal-GPT, InstructBLIP, VisualGLM-6B, PandaGPT, ImageBind-LLM, and LaVIN. Except BLIP-2, other models have been fine-tuned on their instruction tuning datasets.

3.1 Models

All experiments are conducted on NVIDIA V100 GPUs, and we use as large models as possible. **BLIP-2** [30] focuses on the basic pre-training, where the image encoder and the LLM are all frozen and a light-weight Q-Former is trained for multimodal representation alignment and vision-to-language generation. Despite not performing the multimodal instruction tuning, BLIP-2 is still

Model	Poster		Celebrity		Scene		Landmark		Artwork	
	ACC	ACC+	ACC	ACC+	ACC	ACC+	ACC	ACC+	ACC	ACC+
BLIP-2	79.25	62.59	68.53	37.06	81.25	64.00	79.00	59.00	76.50	60.00
MiniGPT-4	37.42	18.37	42.94	22.35	57.25	38.50	46.00	23.00	37.75	18.00
mPLUG-Owl	78.23	57.82	66.18	34.12	78.00	57.50	86.25	73.00	63.25	33.00
ImageBind-LLM	52.72	12.24	55.29	21.18	68.75	44.50	53.00	9.00	54.25	16.50
LLaMA-AdapterV2	65.65	34.01	59.71	26.47	82.50	66.00	82.75	67.50	55.75	14.00
InstructBLIP	74.15	49.66	67.06	34.12	84.00	69.00	59.75	20.00	76.75	57.50
VisualGLM-6B	53.74	12.24	50.88	2.35	81.75	64.50	59.75	24.00	55.25	20.00
Otter	44.90	0.00	50.00	0.00	44.25	0.00	49.50	0.00	41.75	0.00
Multimodal-GPT	42.86	14.97	49.12	24.71	50.50	17.50	48.25	21.50	46.50	13.00
PandaGPT	56.80	19.73	46.47	10.59	72.50	45.50	56.25	13.50	50.25	1.00
LLaVA	50.00	0.00	48.82	0.00	50.00	0.00	50.00	0.00	49.00	0.00
LaVIN	59.18	20.41	37.94	9.41	78.75	58.00	64.00	29.50	59.25	28.00

Table 2: Evaluation results (%) of fine-grained recognition.

Model	Commonsense Reasoning		Numerical Calculation		Text Translation		Code Reasoning	
	ACC	ACC+	ACC	ACC+	ACC	ACC+	ACC	ACC+
BLIP-2	68.57	41.43	40.00	0.00	55.00	10.00	55.00	20.00
MiniGPT-4	45.00	27.14	40.00	15.00	40.00	15.00	65.00	45.00
mPLUG-Owl	57.14	21.43	50.00	10.00	60.00	20.00	47.50	10.00
ImageBind-LLM	40.00	8.57	50.00	5.00	50.00	0.00	50.00	10.00
LLaMA-Adapter V2	58.57	22.86	52.50	10.00	50.00	0.00	50.00	5.00
InstructBLIP	75.00	54.29	35.00	5.00	55.00	10.00	47.50	10.00
VisualGLM-6B	35.00	4.29	45.00	0.00	50.00	0.00	47.50	0.00
Otter	38.57	0.00	20.00	0.00	27.50	0.00	50.00	0.00
Multimodal-GPT	43.57	5.71	42.50	20.00	50.00	10.00	45.00	10.00
PandaGPT	56.43	17.14	50.00	0.00	52.50	5.00	47.50	0.00
LLaVA	47.14	10.00	50.00	0.00	52.50	5.00	50.00	0.00
LaVIN	58.57	28.57	55.00	10.00	47.50	0.00	50.00	0.00

Table 3: Evaluation results (%) of cognition.

able to follow instructions with the LLM's own capabilities. The version under investigation is "blip2-pretrain-flant5xxl" [1].

LLaVA [33] is a pioneer in bringing multimodal instruction tuning to MLLM. The tuning data are generated by language-only GPT-4 through carefully designed prompts. The instructions consist of three types, including conversation, detailed description, and complex reasoning. A projection layer and LLM are updated during training. The tested version is "LLaVA-7B-v0" [6].

MiniGPT-4 [56] performs multimodal instruction tuning on the pre-trained BLIP-2, where only a linear layer is updated. The instructions are mainly based on the image caption task, such as "Describe this image in detail." We use "minigpt4-aligned-with-vicuna13b" [7] for testing.

mPLUG-Owl [50] uses both language-only instruction data and the multimodal instruction data from LLaVA. The visual encoder, the proposed visual abstractor, and LLM (insert a LoRA module [24]) are all updated during training. We test "mplug-owl-llama-7b" [8].

LLaMA-Adapter V2 [20] is only trained on language-only instruction data and image-text pairs via a parameter-efficient manner. Additional expert models, such as detection, can be integrated to boost reasoning. The version of "LLaMA-7B" [5] is used in testing.

InstructBLIP [18] is based on BLIP-2 and reorganizes 26 existing public datasets like image caption and VQA as instruction tuning format. Only the Q-Former is updated during training. "blip2-instruct-flant5xxl" [3] is tested.

VisualGLM-6B [12] is an open-source MLLM but with unclear training details by now. It supports both English and Chinese. We test "VisualGLM-6B" [12].

Otter [29] mixes multimodal in-context learning and multimodal instruction tuning. The framework is based OpenFlamingo [15], whose perceiver module and parts of LLM are updated during training. The instruction data are from VQAv2 [14], GQA [27], LLaVA, and a not public video dataset. The version of "OTTER-9B-LA-InContext" [10] is tested.

Multimodal-GPT [22] is also based on OpenFlamingo and uses both language-only and multimodal instruction data. The latter includes data from LLaVA, Mini-GPT4, and various image caption and

VQA datasets. Only the LoRA module in LLM is updated during training. The tested version is "Multimodal-GPT-9B" [9].

PandaGPT [42] uses multimodal instruction data from LLaVA and MiniGPT-4, and only a linear projection layer as well as the LoRA module in LLM are trainable. The pretrained ImageBind [21] is adopted as the multimodal encoder, which has an attribute of modality alignment. "pandagpt-7b-max-len-512" [11] is used for testing.

ImageBind-LLM [2] is an open-source MLLM and its paper that describes the detailed algorithm is still in preparation. We evaluate "imagebind-LLM-7B" [2].

LaVIN [36] introduces a lightweight adapter as the bridge between vision and LLM. Besides, a routing algorithm is designed to better use language-only and multimodal instruction data. "LAVIN-13B" is tested [4].

3.2 Results

3.2.1 Perception

There are a total of 10 subtasks for the evaluation of the perception ability, from the perspectives of coarse-grained recognition, fine-grained recognition, and OCR. Figs. 2 (3)-(6) show the score leaderboards of individual coarse-grained recognition subtasks. With respect to the object existence, InstructBLIP and LaVIN get the highest score 185, with a 95% accuracy and a 90% accuracy+ listed in Table 1. Contrastively, the second place BLIP-2 and the third place ImageBind-LLM lag behind InstructBLIP by 25 and 56.67 scores, respectively. For the object count, position, and color, InstructBLIP, BLIP-2, and MiniGPT-4 make the top three. What is interesting is that both InstructBLIP and MiniGPT-4 are finetuned on the pre-trained BLIP-2 with instruction data. Note that in the four coarse-grained subtasks, these MLLMs get the worst results on object position, indicating that the current models are not sensitive enough to the position information.

Figs. 2 (7)-(11) show the score leaderboards of individual fine-grained recognition subtasks. Regarding to poster recognition, BLIP-2, mPLUG-Owl, and InstructBLIP are the top three. It is interesting that mPLUG-Owl relatively underperforms in the coarse-grained recognition, but now it exhibits good. This implies that our division of coarse-grained and fine-grained is reasonable and enables us to examine different aspects of MLLMs. For the celebrity recognition, BLIP-2, InstructBLIP, and mPLUG-Owl still take the top three with similar scores. For the scene recognition, InstructBLIP, LLaMA-Adapter V2, and VisualGLM-6B ahead of other MLLMs. This is the first time VisualGLM-6B has broken into the top three. Also noteworthy is PandaGPT, which scores more than a hundred for the first time. For the landmark recognition, top three places are taken by mPLUG-Owl, LLaMA-Adapter V2, and BLIP-2 respectively, of which mPLUG-Owl gets the top spot. For the artwork recognition, BLIP-2, InstructBLIP, and mPLUG-Owl exceed other counterparts, where the first two scores are similar, leading the third by 38+. With respect to OCR listed in Fig. 2 (12), LLaMA-Adapter V2, BLIP-2, and LaVIN get the top three with scores of 125, 110, and 107.5 respectively. As can be seen in Fig. 2 (1), in the leaderboard of the whole perception recognition, BLIP-2, InstructBLIP, and LLaMA-Adapter V2 come in top three, closely followed by mPLUG-Owl and LaVIN. Their perception scores are 1293.84, 1212.82, 972.67, 967.35, and 963.61, respectively.

3.2.2 Cognition

There are four subtasks for the evaluation of the cognition ability, including commonsense reasoning, numerical calculation, text translation, and code reasoning. Figs. 2 (13)-(16) plot the score leaderboards of individual subtasks. In terms of the commonsense reasoning, the "ever-victorious generals" InstructBLIP and BLIP-2 still exceed other MLLMs, especially InstructBLIP, which gets a score of 129.29. With respect to numerical calculation and text translation, although the designed questions are moderate in difficulty, these MLLMs do not perform satisfactorily, with no score beyond 80. This suggests that MLLMs have a lot of room for improvement in these capabilities. For the code reasoning, MiniGPT-4 achieves a high score of 110, far ahead of other counterparts. This is reminiscent of the example it showed of writing website codes from an image. For all of the cognition tasks, MiniGPT-4, InstructBLIP, and BLIP-2 win the gold, silver, and bronze medals respectively, as shown in Fig. 2 (2).

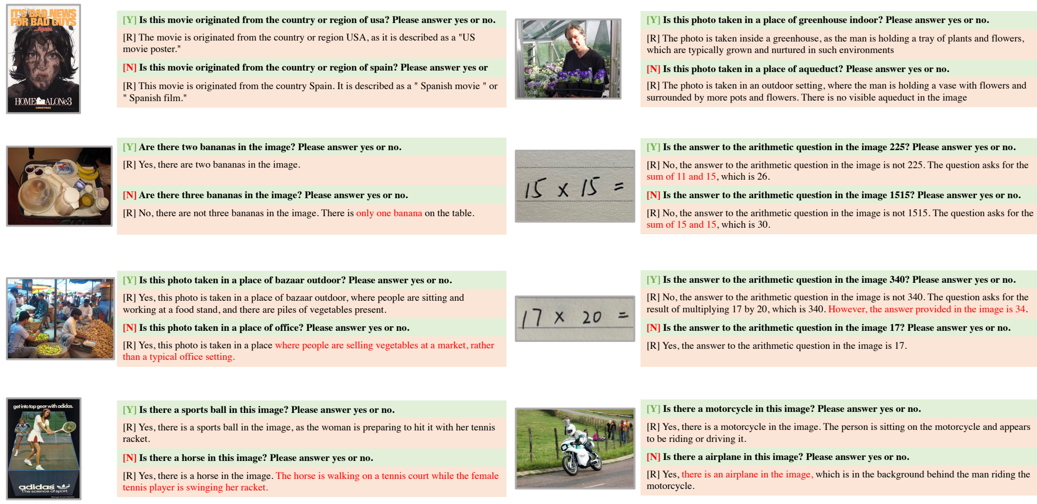

Figure 4: Common problems revealed in our experiments. [Y]/[N] means the ground truth answer is yes/no. [R] is the generated answer of MLLM.

4 Analysis

We conclude four common problems that largely affect the performance of MLLMs. **The first problem is not following instructions.** Although we have adopted a very concise instruction design, there are MLLMs that answer freely rather than following instructions. For example, as shown in the first row of Fig. 4, the instruction has claimed "Please answer yes or no", but the MLLM only makes a declarative expression. If no "yes" or "no" is appeared at the beginning of the generated languages, the model is judged to make a wrong answer. We argue that a good MLLM (especially after instruction tuning) should be able to follow such a simple instruction, which is also very common in everyday life.

The second problem is a lack of perception. As shown in the second row of Fig. 4, the MLLM misidentifies the number of bananas in the first image, and misreads the characters in the second image, resulting in wrong answers. We notice that the performance of perception is vulnerable to the nuance of instructions, since the two instructions of the same image differ in only one word, but lead to completely different and even contradictory perception results.

The third problem is a lack of reasoning. In the third row of Fig. 4, we can see from the red text that the MLLM already knows that the first image is not an office place, but still gives an incorrect answer of "yes". Analogously, in the second image, the MLLM has calculated the right arithmetic result, but finally delivers a wrong answer. These phenomena indicate that the logic chain is broken during the reasoning process of MLLMs. Adding CoT prompts, such as "Let's think step by step" [19], may yield better results. We look forward to a further in-depth research.

The fourth problem is object hallucination following instructions, which is exemplified in the fourth row of Fig. 4. When the instruction contains descriptions of an object that does not appear in the image, the MLLM will imagine that the object exists and ultimately gives a "yes" answer. Such a case of constantly answering "yes" results in an accuracy about 50% and an accuracy+ about 0, as shown in Tables 1, 2, and 3. This suggests an urgent need to suppress hallucinations, and the community should take into account of the reliability of the generated answers.

5 Conclusion

This paper has presented the first MLLM evaluation benchmark MME that has four distinct characteristics in terms of task type, data source, instruction design, quantitative statistics. We evaluate 12 advanced MLLMs on MME and the experimental results show that there is still a large room to improve. We also summarize the common problem raised in experimental results, providing valuable guidance for the development of MLLM.

References

[1] Blip-2. https://github.com/salesforce/LAVIS/tree/main/projects/blip2, 2023.

[2] Imagebind-llm. https://github.com/OpenGVLab/LLaMA-Adapter/tree/main/imagebind_LLM, 2023.

[3] Instructblip. https://github.com/salesforce/LAVIS/tree/main/projects/instructblip, 2023.

[4] Lavin. https://github.com/luogen1996/LaVIN, 2023.

[5] Llama-adapter v2. https://github.com/OpenGVLab/LLaMA-Adapter, 2023.

[6] Llava. https://github.com/haotian-liu/LLaVA, 2023.

[7] Minigpt-4. https://github.com/Vision-CAIR/MiniGPT-4, 2023.

[8] mplug-owl. https://huggingface.co/MAGAer13/mplug-owl-llama-7b, 2023.

[9] Multimodal-gpt. https://github.com/open-mmlab/Multimodal-GPT, 2023.

[10] Otter. https://github.com/Luodian, 2023.

[11] Pandagpt. https://github.com/yxuansu/PandaGPT, 2023.

[12] Visualglm-6b. https://github.com/THUDM/VisualGLM-6B, 2023.

[13] 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. *NeurIPS*, 2022.

[14] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *ICCV*, 2015.

[15] Anas Awadalla, Irena Gao, Joshua Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Jenia Jitsev, Simon Kornblith, Pang Wei Koh, Gabriel Ilharco, Mitchell Wortsman, and Ludwig Schmidt. Openflamingo, 2023. URL https://doi.org/10.5281/zenodo.7733589.

[16] 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. *NeurIPS*, 2020.

[17] 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:1504.00325*, 2015.

[18] 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:2305.06500*, 2023.

[19] Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, et al. Palm-e: An embodied multimodal language model. *arXiv preprint:2303.03378*, 2023.

[20] Peng Gao, Jiaming Han, Renrui Zhang, Ziyi Lin, Shijie Geng, Aojun Zhou, Wei Zhang, Pan Lu, Conghui He, Xiangyu Yue, et al. Llama-adapter v2: Parameter-efficient visual instruction model. *arXiv preprint:2304.15010*, 2023.

[21] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind: One embedding space to bind them all. In *CVPR*, 2023.

[22] 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. *arXiv preprint:2305.04790*, 2023.

[23] 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 *CVPR*, 2017.

[24] 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:2106.09685*, 2021.

[25] Qingqiu Huang, Yu Xiong, Anyi Rao, Jiaze Wang, and Dahua Lin. Movienet: A holistic dataset for movie understanding. In *ECCV*, 2020.

[26] Shaohan Huang, Li Dong, Wenhui Wang, Yaru Hao, Saksham Singhal, Shuming Ma, Tengchao Lv, Lei Cui, Owais Khan Mohammed, Qiang Liu, et al. Language is not all you need: Aligning perception with language models. *arXiv preprint:2302.14045*, 2023.

[27] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *CVPR*, 2019.

[28] Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Fanyi Pu, Jingkang Yang, Chunyuan Li, and Ziwei Liu. Mimic-it: Multi-modal in-context instruction tuning. *arXiv preprint:2306.05425*, 2023.

[29] Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Jingkang Yang, and Ziwei Liu. Otter: A multi-modal model with in-context instruction tuning. *arXiv preprint:2305.03726*, 2023.

[30] 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:2301.12597*, 2023.

[31] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. *arXiv preprint:2305.10355*, 2023.

[32] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014.

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

[34] Yuliang Liu, Lianwen Jin, Shuaitao Zhang, Canjie Luo, and Sheng Zhang. Curved scene text detection via transverse and longitudinal sequence connection. *PR*, 2019.

[35] 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. *NeurIPS*, 2022.

[36] Gen Luo, Yiyi Zhou, Tianhe Ren, Shengxin Chen, Xiaoshuai Sun, and Rongrong Ji. Cheap and quick: Efficient vision-language instruction tuning for large language models. *arXiv preprint:2305.15023*, 2023.

[37] Hui Mao, Ming Cheung, and James She. Deepart: Learning joint representations of visual arts. In *ICM*, 2017.

[38] Hui Mao, James She, and Ming Cheung. Visual arts search on mobile devices. *TOMM*, 2019.

[39] Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *CVPR*, 2019.

[40] OpenAI. Gpt-4 technical report. *arXiv preprint:2303.08774*, 2023.

[41] Yongliang Shen, Kaitao Song, Xu Tan, Dongsheng Li, Weiming Lu, and Yueting Zhuang. Hugginggpt: Solving ai tasks with chatgpt and its friends in huggingface. *arXiv preprint:2303.17580*, 2023.

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

[43] 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:2302.13971*, 2023.

[44] Wenhai Wang, Zhe Chen, Xiaokang Chen, Jiannan Wu, Xizhou Zhu, Gang Zeng, Ping Luo, Tong Lu, Jie Zhou, Yu Qiao, et al. Visionllm: Large language model is also an open-ended decoder for vision-centric tasks. *arXiv preprint:2305.11175*, 2023.

[45] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny Zhou. Chain of thought prompting elicits reasoning in large language models. *arXiv preprint:2201.11903*, 2022.

[46] Tobias Weyand, Andre Araujo, Bingyi Cao, and Jack Sim. Google landmarks dataset v2-a large-scale benchmark for instance-level recognition and retrieval. In *CVPR*, 2020.

[47] Chenfei Wu, Shengming Yin, Weizhen Qi, Xiaodong Wang, Zecheng Tang, and Nan Duan. Visual chatgpt: Talking, drawing and editing with visual foundation models. *arXiv preprint:2303.04671*, 2023.

[48] Peng Xu, Wenqi Shao, Kaipeng Zhang, Peng Gao, Shuo Liu, Meng Lei, Fanqing Meng, Siyuan Huang, Yu Qiao, and Ping Luo. Lvlm-ehub: A comprehensive evaluation benchmark for large vision-language models. *arXiv preprint:2306.09265*, 2023.

[49] Zhiyang Xu, Ying Shen, and Lifu Huang. Multiinstruct: Improving multi-modal zero-shot learning via instruction tuning. *arXiv preprint:2212.10773*, 2022.

[50] 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:2304.14178*, 2023.

[51] Zhenfei Yin, Jiong Wang, Jianjian Cao, Zhelun Shi, Dingning Liu, Mukai Li, Lu Sheng, Lei Bai, Xiaoshui Huang, Zhiyong Wang, et al. Lamm: Language-assisted multi-modal instruction-tuning dataset, framework, and benchmark. *arXiv preprint:2306.06687*, 2023.

[52] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. A survey of large language models. *arXiv preprint:2303.18223*, 2023.

[53] Yunqing Zhao, Tianyu Pang, Chao Du, Xiao Yang, Chongxuan Li, Ngai-Man Cheung, and Min Lin. On evaluating adversarial robustness of large vision-language models. *arXiv preprint:2305.16934*, 2023.

[54] Zijia Zhao, Longteng Guo, Tongtian Yue, Sihan Chen, Shuai Shao, Xinxin Zhu, Zehuan Yuan, and Jing Liu. Chatbridge: Bridging modalities with large language model as a language catalyst. *arXiv preprint:2305.16103*, 2023.

[55] Bolei Zhou, Agata Lapedriza, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. Learning deep features for scene recognition using places database. *NeurIPS*, 2014.

[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:2304.10592*, 2023.