MME: A Comprehensive Evaluation Benchmark for Multimodal Large Language Models

Chaoyou Fu[1,2,♠], Peixian Chen[3], Yunhang Shen[3], Yulei Qin[3], Mengdan Zhang[3]
Xu Lin[3], Jinrui Yang[3], Xiawu Zheng[4], Ke Li[3,†], Xing Sun[3]
Yunsheng Wu[3], Rongrong Ji[4], Caifeng Shan[1,2], Ran He[5]

[1]State Key Laboratory for Novel Software Technology, Nanjing University
[2]School of Intelligence Science and Technology, Nanjing University
[3]Tencent Youtu Lab [4]Xiamen University [5]CASIA
♠ Project Leader † Corresponding Author

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 comprehensive **M**LLM **E**valuation benchmark **MME**. 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 30 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. The data are released at the project page: `https://github.com/BradyFU/Awesome-Multimodal-Large-Language-Models/tree/Evaluation`.

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) [39, 21, 25, 13, 42, 44]. It refers to using LLM as a brain to process multimodal information and give reasoning results [53]. Equipped with the powerful LLM, MLLM is expected to address more complex multi-modal tasks [13, 48, 40, 28, 60, 61, 32, 31, 59, 55, 14].The three representative abilities of LLM [62], including instruction following [43], In-Context Learning (ICL) [10], and Chain-of-Thought (CoT) [47] are also manifested in multimodality. For example, Flamingo [8] turns on multimodal ICL, which can adapt to new tasks by giving a few examples. PaLM-E [13] achieves amazing OCR-free math reasoning via CoT. GPT-4V [39] shows even more ability in a variety of complex reasoning tasks [50]. MiniGPT-4 [66] implements GPT-4[39]-like instruction following capabilities, such as converting images into corresponding website codes, by introducing multimodal instruction tuning. These emergent abilities of MLLMs 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

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Track on Datasets and Benchmarks.

 https://doi.org/10.52202/085713-4899

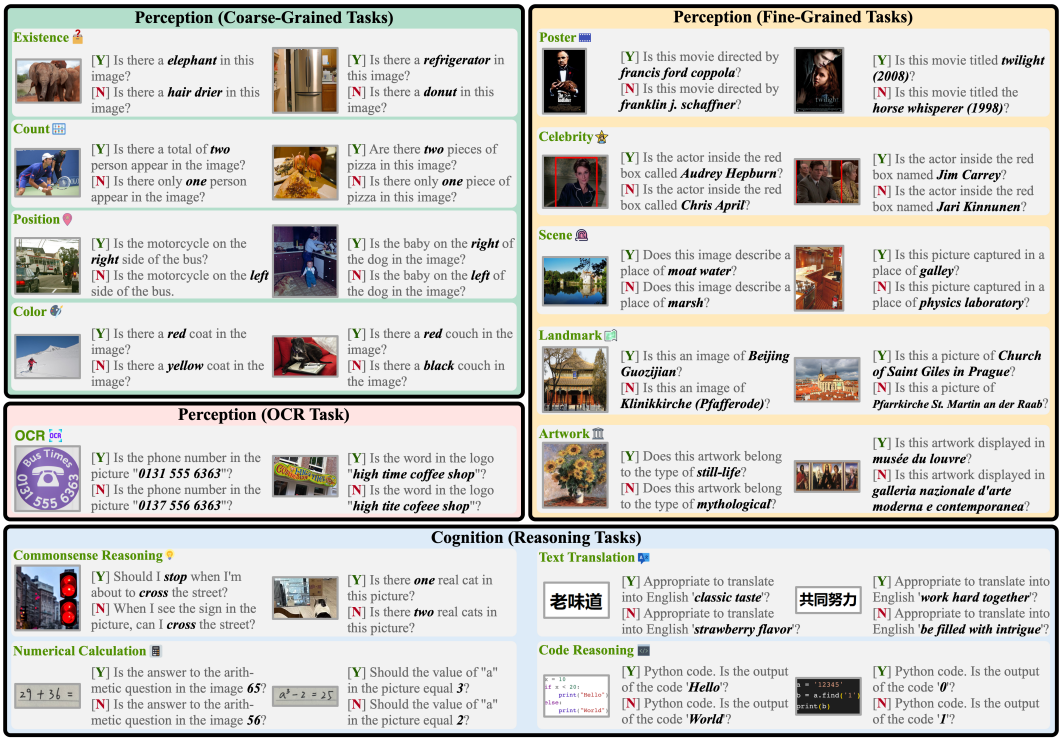

Figure 1: Diagram of our MME benchmark. It evaluates MLLMs from both perception and cognition, including a total of 14 subtasks. Each image corresponds to two questions whose answers are marked yes [Y] and no [N], respectively. The instruction consists of a question followed by "Please answer yes or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

three common quantitative evaluation manners for MLLMs have their limitations that are difficult to comprehensively evaluate performance. Specifically, the first manner [51, 12, 46] evaluates on existing traditional multimodal datasets, such as image caption [11] and VQA [17, 38, 34]. However, on the one hand, it may be hard to reflect the emergent abilities of MLLMs on these datasets. On the 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 [52] is to collect data for an open-ended evaluation, but either the data is unavailable to public by now [64] or the amount is small (only 50 images) [52]. The third manner focuses on one aspect of MLLMs, such as object hallucination [26] or adversarial robustness [63], 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 MLLMs to the instructions should be intuitive and convenient for quantitative analysis. The open-ended answer of MLLMs poses significant challenges to the quantization. Existing methods tend to use GPT or manual scoring [22, 30, 52], 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:

Rank	Model	Score
1	WeMM	1621.66
2	InfMLLM	1567.99
3	SPHINX	1560.15
4	Lion	1545.80
5	LLaVA	1531.31
6	XComposer-VL	1528.45
7	Qwen-VL-Chat	1487.58
8	mPLUG-Owl2	1450.20
9	Skywork-MM	1419.08
10	GPT-4V	1409.43

(1) **Perception**

Rank	Model	Score
1	GPT-4V	517.14
2	Lion	445.71
3	WeMM	445.00
4	MMICL	428.93
5	XComposer-VL	391.07
6	Qwen-VL-Chat	360.71
7	LLaMA-Adapter V2	356.43
8	Skywork-MM	356.43
9	InfMLLM	347.14
10	BLIVA	331.43

(2) **Cognition**

Rank	Model	Score
1	Otter	195.00
1	Lynx	195.00
1	WeMM	195.00
1	Muffin	195.00
1	SPHINX	195.00
2	GIT2	190.00
2	XComposer-VL	190.00
2	Lion	190.00
2	GPT-4V	190.00
2	InfMLLM	190.00

(3) Existence

Rank	Model	Score
1	Muffin	163.33
2	MMICL	160.00
2	GPT-4V	160.00
2	SPHINX	160.00
3	XComposer-VL	158.33
4	LLaVA	155.00
4	Lion	155.00
4	mPLUG-Owl2	155.00
5	Lynx	151.67
5	Skywork-MM	151.67

(4) Count

Rank	Model	Score
1	Lion	153.33
1	SPHINX	153.33
2	InfMLLM	143.33
3	LLaVA	133.33
4	Qwen-VL-Chat	128.33
5	WeMM	126.67
5	XComposer-VL	126.67
6	GIT2	96.67
7	GPT-4V	95.00
8	Lynx	90.00

(5) Position

Rank	Model	Score
1	InfMLLM	185.00
2	BLIVA	180.00
2	Lion	180.00
3	LLaVA	170.00
3	Lynx	170.00
3	Qwen-VL-Chat	170.00
4	WeMM	168.33
5	LRV-Instruction	165.00
5	XComposer-VL	165.00
5	Muffin	165.00

(6) Color

Rank	Model	Score
1	GPT-4V	192.18
2	Lion	181.63
3	Qwen-VL-Chat	178.57
4	Skywork-MM	175.85
5	SPHINX	164.29
6	InfMLLM	163.27
7	XComposer-VL	161.90
8	LLaVA	160.54
8	WeMM	160.54
9	mPLUG-Owl2	160.20

(7) Poster

Rank	Model	Score
1	WeMM	179.12
2	SPHINX	177.94
3	Otter	172.65
4	mPLUG-Owl2	164.41
5	Cheetor	164.12
6	InfMLLM	161.47
7	Skywork-MM	160.29
8	LLaVA	152.94
9	Lion	150.59
10	XComposer-VL	150.29

(8) Celebrity

Rank	Model	Score
1	WeMM	176.25
2	InfMLLM	165.25
3	Lynx	164.50
4	LLaVA	161.25
5	SPHINX	160.00
6	XComposer-VL	159.75
7	Lion	159.00
8	Otter	158.75
9	GIT2	158.50
10	Octopus	157.25

(9) Scene

Rank	Model	Score
1	Lion	173.00
2	WeMM	172.25
3	LLaVA	170.50
4	SPHINX	168.09
5	LLaMA-Adapter V2	167.84
6	InfMLLM	167.00
7	XComposer-VL	165.25
8	Qwen-VL-Chat	164.00
9	Lynx	162.00
10	LRV-Instruction	160.53

(10) Landmark

Rank	Model	Score
1	WeMM	156.00
2	GPT-4V	148.00
3	GIT2	146.25
4	BLIP-2	136.50
5	MMICL	135.50
6	InstructBLIP	134.25
6	mPLUG-Owl2	134.25
7	SPHINX	134.00
8	BLIVA	133.25
9	Lion	130.75

(11) Artwork

Rank	Model	Score
1	GPT-4V	185.00
2	Skywork-MM	162.50
3	WeMM	147.50
4	Qwen-VL-Chat	140.00
5	InfMLLM	132.50
6	LLaVA	125.00
6	XComposer-VL	125.00
7	BLIP-2	110.00
7	LRV-Instruction	110.00
8	LaVIN	107.50

(12) OCR

Rank	Model	Score
1	GPT-4V	142.14
2	WeMM	140.00
3	XComposer-VL	138.57
4	BLIVA	136.43
4	MMICL	136.43
5	InfMLLM	132.14
6	Qwen-VL-Chat	130.71
7	SPHINX	130.00
8	InstructBLIP	129.29
9	LLaVA	127.86

(13) Commonsense Reasoning

Rank	Model	Score
1	GPT-4V	130.00
2	Lion	105.00
3	Skywork-MM	95.00
4	MMICL	82.50
5	Cheetor	77.50
6	Otter	72.50
7	LRV-Instruction	70.00
8	LaVIN	65.00
9	Multimodal-GPT	62.50
10	mPLUG-Owl	60.00

(14) Numerical Calculation

Rank	Model	Score
1	Qwen-VL-Chat	147.50
1	Lion	147.50
2	MMICL	132.50
3	WeMM	130.00
4	LLaMA-Adapter V2	112.50
4	XComposer-VL	112.50
5	Octopus	102.50
5	mPLUG-Owl2	102.50
5	InfMLLM	102.50
6	LRV-Instruction	85.00

(15) Text Translation

Rank	Model	Score
1	GPT-4V	170.00
2	WeMM	117.50
3	LLaMA-Adapter V2	90.00
4	Cheetor	87.50
5	XComposer-VL	85.00
6	MMICL	77.50
7	BLIP-2	75.00
8	LRV-Instruction	72.50
9	Otter	70.00
10	Lion	67.50

(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 and 2. A total of 30 advanced MLLMs joint the leaderboards. For the sake of presentation, we only show 10 models for each list, in which the top three ones are given clear trophy logos.

- 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.

https://doi.org/10.52202/085713-4899

Model	Existence		Count		Position		Color		OCR		Poseter		Celebrity	
	ACC	ACC+	ACC	ACC+	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	79.25	62.59	68.53	37.06
mPLUG-Owl	73.33	46.67	50.00	0.00	50.00	0.00	51.67	3.33	55.00	10.00	78.23	57.82	66.18	34.12
ImageBind-LLM	75.00	53.33	50.00	10.00	43.33	3.33	56.67	16.67	60.00	20.00	52.72	12.24	55.29	21.18
InstructBLIP	95.00	90.00	80.00	63.33	53.33	13.33	83.33	70.00	57.50	15.00	74.15	49.66	67.06	34.12
VisualGLM-6B	61.67	23.33	50.00	0.00	48.33	0.00	51.67	3.33	42.50	0.00	53.74	12.24	50.88	2.35
Multimodal-GPT	48.33	13.33	48.33	6.67	45.00	13.33	55.00	13.33	57.50	25.00	42.86	14.97	49.12	24.71
PandaGPT	56.67	13.33	50.00	0.00	50.00	0.00	50.00	0.00	50.00	0.00	56.80	19.73	46.47	10.59
LaVIN	95.00	90.00	61.67	26.67	53.33	10.00	58.33	16.67	67.50	40.00	59.18	20.41	37.94	9.41
Cheetor	93.33	86.67	63.33	33.33	56.67	23.33	70.00	46.67	65.00	35.00	81.29	65.99	87.65	76.47
GIT2	96.67	93.33	71.67	46.67	60.00	36.67	85.00	73.33	55.00	10.00	61.56	51.02	81.18	64.71
GPT-4V	96.67	93.33	**86.67**	73.33	65.00	30.00	80.00	70.00	**95.00**	**90.00**	**96.94**	**95.24**	0.00	0.00
XComposer-VL	96.67	93.33	85.00	73.33	73.33	53.33	88.33	76.67	75.00	50.00	85.71	76.19	83.24	67.06
LLaVA	95.00	90.00	85.00	70.00	76.67	56.67	90.00	80.00	75.00	50.00	86.39	74.15	83.53	69.41
LRV-Instruction	88.33	76.67	68.33	43.33	56.67	30.00	88.33	76.67	70.00	40.00	78.77	60.27	67.35	45.29
Lion	96.67	93.33	85.00	70.00	**83.33**	**70.00**	93.33	86.67	57.50	15.00	93.88	87.76	82.94	67.65
Lynx	**98.33**	**96.67**	81.67	70.00	60.00	30.00	90.00	80.00	57.50	20.00	74.49	50.34	71.76	46.47
MMICL	90.00	80.00	**86.67**	73.33	55.00	26.67	83.33	73.33	60.00	40.00	81.63	64.63	79.41	62.35
Muffin	**98.33**	**96.67**	**86.67**	**76.67**	53.33	13.33	88.33	76.67	52.50	5.00	78.57	59.18	56.47	25.29
Octopus	93.33	86.67	50.00	3.33	45.00	3.33	66.67	36.67	55.00	10.00	78.23	59.86	75.29	54.12
Otter	**98.33**	**96.67**	58.33	30.00	60.00	26.67	70.00	43.33	57.50	15.00	78.91	59.86	90.88	81.76
Qwen-VL-Chat	85.00	73.33	83.33	66.67	75.00	53.33	90.00	80.00	80.00	60.00	92.18	86.39	72.35	48.24
SPHINX	**98.33**	**96.67**	**86.67**	73.33	**83.33**	**70.00**	86.67	73.33	62.50	25.00	87.41	76.87	**92.65**	85.29
Skywork-MM	93.33	86.67	81.67	70.00	46.67	16.67	81.67	63.33	87.50	75.00	91.50	84.35	86.76	73.53
VPGTrans	56.67	13.33	61.67	23.33	53.33	10.00	56.67	16.67	57.50	20.00	60.88	23.13	51.18	2.35
WeMM	**98.33**	**96.67**	80.00	60.00	73.33	53.33	88.33	80.00	82.50	65.00	86.39	74.15	**92.65**	**86.47**
BLIVA	93.33	86.67	78.33	60.00	58.33	23.33	93.33	86.67	62.50	25.00	84.35	70.75	80.29	60.59
InfMLLM	96.67	93.33	81.67	70.00	80.00	63.33	**95.00**	**90.00**	77.50	55.00	87.07	76.19	86.18	75.29
LLaMA-AdapterV2	95.00	90.00	73.33	60.00	50.00	6.67	71.67	46.67	67.50	35.00	81.97	65.99	77.94	58.82
MiniGPT-4	51.67	16.67	48.33	6.67	43.33	0.00	55.00	20.00	52.50	5.00	36.39	5.44	45.00	9.41
mPLUG-Owl2	95.00	90.00	85.00	70.00	61.67	26.67	83.33	66.67	67.50	35.00	86.73	73.47	87.94	76.47

Table 1: Evaluation results on the subtasks of existence, count, position, color, OCR, poster, and celebrity. The top two results on each subtask are **bolded** and underlined, respectively.

- 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 30 advanced MLLMs on the 14 subtasks. The evaluated MLLMs include BLIP-2 [25], InstructBLIP [12], MiniGPT-4 [66], PandaGPT [41], Multimodal-GPT [16], VisualGLM-6B [5], ImageBind-LLM [18], VPGTrans [58], LaVIN [35], mPLUG-Owl [52], Octopus [3], Muffin [56], Otter [23], LRV-Instruction [29], Cheetor [24], LLaMA-Adapter-v2 [15], GIT2 [45], BLIVA [19], Lynx [57], MMICL [61], GPT-4V [39], Skywork-MM [4], mPLUG-Owl2 [52], Qwen-VL-Chat [9], XComposer-VL [7], LLaVA [30], Lion [2], SPHINX [28], InfMLLM [1], and WeMM [6]. 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. We can see the range that current MLLMs can reach in each capability dimension. 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 [54, 26], as shown in Fig. 4. It is expected that 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. (2) A total of 30 up-to-date MLLMs are evaluated on our MME. (3) We summarize the exposed problems in experiments, proving guidance for the evolution of MLLMs.

Model	Scene		Landmark		Artwork		Comm.		Num.		Text.		Code.	
	ACC	ACC+	ACC	ACC+	ACC	ACC+	ACC	ACC+	ACC	ACC+	ACC	ACC+	ACC	ACC+
BLIP-2	81.25	64.00	79.00	59.00	76.50	60.00	68.57	41.43	40.00	0.00	55.00	10.00	55.00	20.00
mPLUG-Owl	78.00	57.50	86.25	73.00	63.25	33.00	57.14	21.43	50.00	10.00	60.00	20.00	47.50	10.00
ImageBind-LLM	68.75	44.50	53.00	9.00	54.25	16.50	40.00	8.57	50.00	5.00	50.00	0.00	50.00	10.00
InstructBLIP	84.00	69.00	59.75	20.00	76.75	57.50	75.00	54.29	35.00	5.00	55.00	10.00	47.50	10.00
VisualGLM-6B	81.75	64.50	59.75	24.00	55.25	20.00	35.00	4.29	45.00	0.00	50.00	0.00	47.50	0.00
Multimodal-GPT	50.50	17.50	48.25	21.50	46.50	13.00	43.57	5.71	42.50	20.00	50.00	10.00	45.00	10.00
PandaGPT	72.50	45.50	56.25	13.50	50.25	1.00	56.43	17.14	50.00	0.00	52.50	5.00	47.50	0.00
LaVIN	78.75	58.00	64.00	29.50	59.25	28.00	58.57	28.57	55.00	10.00	47.50	0.00	50.00	0.00
Cheetor	84.50	71.50	81.41	64.32	67.50	46.00	64.29	34.29	57.50	20.00	42.50	15.00	57.50	30.00
GIT2	86.00	72.50	79.50	61.00	81.25	65.00	66.43	32.86	40.00	10.00	52.50	15.00	45.00	0.00
GPT-4V	83.50	67.50	79.25	59.00	82.00	66.00	79.29	**62.86**	**75.00**	**55.00**	55.00	20.00	**90.00**	**80.00**
XComposer-VL	86.25	73.50	87.75	77.50	73.25	53.00	77.14	61.43	40.00	15.00	67.50	45.00	55.00	30.00
LLaVA	86.75	74.50	90.00	80.50	70.75	47.00	73.57	54.29	37.50	5.00	57.50	20.00	42.50	5.00
LRV-Instruction	81.04	65.93	86.84	73.68	63.75	37.50	65.00	35.71	45.00	25.00	55.00	30.00	57.50	15.00
Lion	85.50	73.50	**91.00**	**82.00**	75.75	55.00	74.29	51.43	65.00	40.00	**82.50**	**65.00**	42.50	25.00
Lynx	88.00	76.50	87.00	75.00	72.50	47.00	66.43	44.29	17.50	0.00	42.50	0.00	45.00	0.00
MMICL	83.75	70.00	76.96	59.16	76.50	59.00	76.43	60.00	47.50	35.00	72.50	60.00	47.50	30.00
Muffin	83.75	67.50	81.25	65.00	71.00	45.50	68.57	41.43	45.00	0.00	57.50	15.00	47.50	15.00
Octopus	84.75	72.50	75.00	51.00	64.50	30.50	64.29	35.71	47.50	0.00	62.50	40.00	47.50	15.00
Otter	86.25	72.50	78.75	58.50	75.00	54.00	66.43	40.00	52.50	20.00	52.50	5.00	55.00	15.00
Qwen-VL-Chat	83.75	68.50	88.00	76.00	73.50	52.00	76.43	54.29	35.00	5.00	**82.50**	**65.00**	32.50	10.00
SPHINX	86.50	73.50	89.20	78.89	77.00	57.00	75.71	54.29	40.00	15.00	50.00	25.00	50.00	0.00
Skywork-MM	79.29	59.60	75.51	51.53	69.43	45.08	72.14	54.29	65.00	30.00	60.00	20.00	45.00	10.00
VPGTrans	80.25	61.50	54.75	10.00	57.75	19.50	50.00	14.29	50.00	0.00	57.50	20.00	52.50	5.00
WeMM	**91.75**	**84.50**	90.75	81.50	**85.00**	**71.00**	**80.00**	60.00	47.50	10.00	75.00	55.00	72.50	45.00
BLIVA	83.50	68.00	63.00	26.50	77.25	56.00	77.86	58.57	47.50	10.00	57.50	20.00	50.00	10.00
InfMLLM	87.75	77.50	88.50	78.50	68.00	40.50	76.43	55.71	45.00	15.00	67.50	35.00	42.50	10.00
LLaMA-AdapterV2	85.25	71.00	88.44	79.40	73.25	50.50	67.86	38.57	47.50	0.00	67.50	45.00	60.00	30.00
MiniGPT-4	54.25	17.50	45.50	8.50	50.00	10.50	46.43	12.86	45.00	0.00	0.00	0.00	40.00	0.00
mPLUG-Owl2	83.25	70.00	85.75	71.50	77.25	57.00	71.43	44.29	35.00	0.00	67.50	35.00	45.00	15.00

Table 2: Evaluation results on the subtasks of scene, landmark, artwork, commonsense reasoning, numerical calculation, text translation, and code reasoning. The top two results on each subtask are **bolded** and underlined, respectively.

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 the two questions correctly, it appears more confident that the MLLM actually comprehends the image and the corresponding knowledge behind it, rather than just guessing.

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. Therefore, the full scores of perception and cognition are 2000 and 800, respectively.

 https://doi.org/10.52202/085713-4899

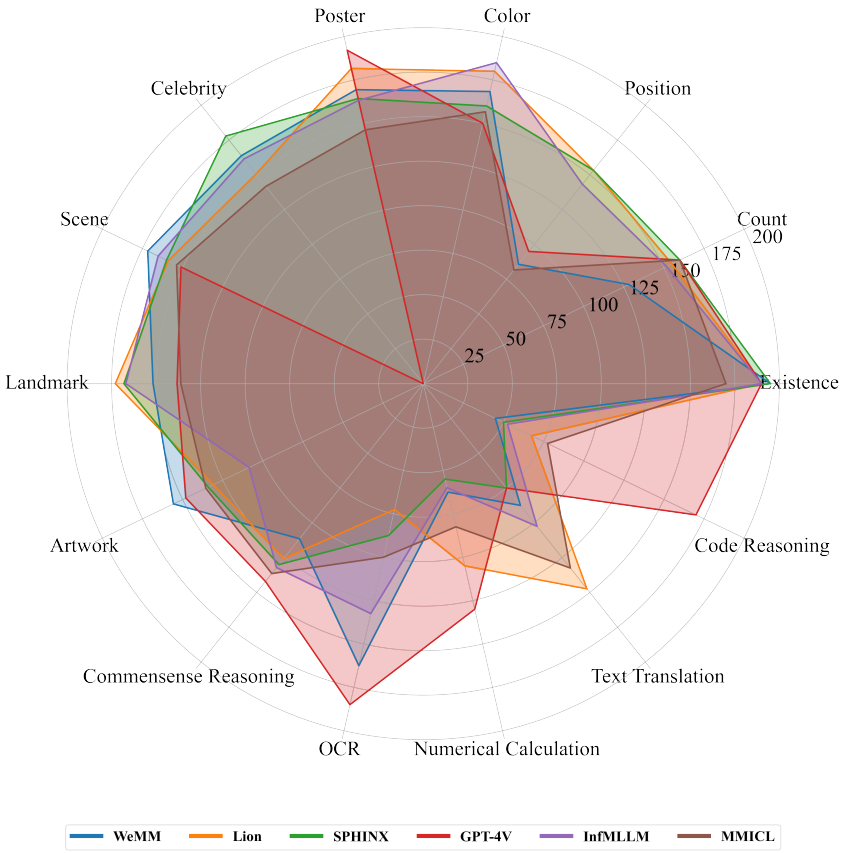

Figure 3: Comparison of 6 top MLLMs on 14 subtasks. The full score of each subtask is 200.

2.3 Data Collection

2.3.1 Perception Tasks

We argue that perception is one of the most fundamental capabilities of MLLMs, and the lack of perception will easily lead to the object hallucination problem [54, 26]. That is, MLLM will answer 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 [27], 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 MLLMs 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 MLLMs. 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 [20, 36, 37, 65, 49] and all of the instructions are manually designed.

OCR. Optical Character Recognition (OCR) is also a foundational capability of MLLMs, serving for subsequent text-based tasks such as text translation and text understanding. The images are sampled from [33] and all of the instruction-answer pairs are manually designed. Considering that MLLMs

are 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 MLLMs over previous traditional methods. In order to infer the correct answer, MLLMs need to follow the instruction, perceive the contents of the image, and invoke the knowledge reserved in LLMs, 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 [34] that requires specialized knowledge, the commonsense refers to the basic knowledge in daily life. For example, given a photo of a down jacket, asking MLLMs 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 MLLMs 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 MLLMs 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 [21]. In this version, we only consider relatively easy arithmetic problems, such as addition and multiplication. There are 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 [5] supports both English and Chinese, we set the text translation subtask. It requires MLLMs 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 MLLMs 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 MLLMs 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 [66]. 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, a total of 30 MLLMs are evaluated on our MME benchmark, including BLIP-2 [25], InstructBLIP [12], MiniGPT-4 [66], PandaGPT [41], Multimodal-GPT [16], VisualGLM-6B [5], ImageBind-LLM [18], VPGTrans [58], LaVIN [35], mPLUG-Owl [52], Octopus [3], Muffin [56], Otter [23], LRV-Instruction [29], Cheetor [24], LLaMA-Adapter-v2 [15], GIT2 [45], BLIVA [19], Lynx [57], MMICL [61], GPT-4V [39], Skywork-MM [4], mPLUG-Owl2 [52], Qwen-VL-Chat [9], XComposer-VL [7], LLaVA [30], Lion [2], SPHINX [28], InfMLLM [1], and WeMM [6].

3.1 Results

3.1.1 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, Otter, Lynx, WeMM, Muffin, and SPHINX get the highest score 195, with a 98.33% accuracy and a 96.67% accuracy+ listed in Table 1. Contrastively, the second place, including GIT2, XComposer-VL, Lion, GPT-4V, and etc, lag behind the first place only by 5 scores. The results show that these models already have a good performance on object existence. For the object count, position, and color, Muffin, Lion (parallel with SPHINX), and InfMLLM make the top one, respectively. It suggests that different models have their own strengths. Note that in the four coarse-grained subtasks, these

 https://doi.org/10.52202/085713-4899

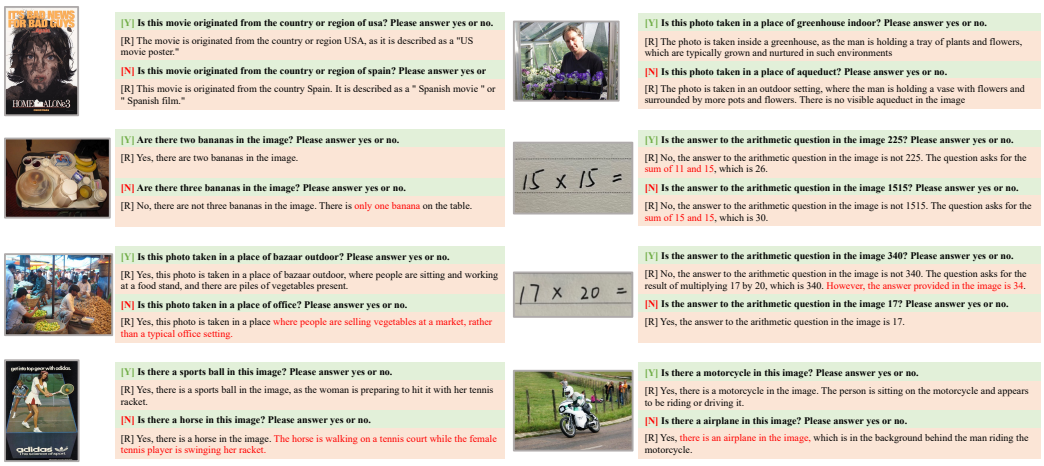

Figure 4: Common problems revealed in experiments. [Y]/[N] means the ground truth answer is yes/no. [R] is the generated answer.

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) display the score leaderboards of individual fine-grained recognition subtasks. Regarding to poster recognition, GPT-4V, Lion, and Qwen-VL-Chat are the top three. It is interesting that Qwen-VL-Chat 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, enabling us to examine different aspects of MLLMs. For the celebrity recognition, WeMM, SPHINX, and Otter take the top three with similar scores. It is worth noting that GPT-4V refuses to answer questions that involve individuals, resulting in a zero score in the celebrity subtask. For the scene recognition, WeMM, InfMLLM, and Lynx ahead of other MLLMs. This is the first time InfMLLM and Lynx have broken into the top three in the fine-grained recognition subtasks. For the landmark recognition, top three places are taken by Lion, WeMM, and LLaVA respectively, of which Lion gets the top spot. For the artwork recognition, WeMM, GPT-4V, and GIT2 exceed other counterparts, where the last two scores are similar. Note that GPT-4V declines to answer some questions about private art collection, which lowers its score. With respect to OCR listed in Fig. 2 (12), GPT-4V, Skywork-MM, and WeMM get the top three with scores of 185, 162.5, and 147.5 respectively. GPT-4V presents a huge advantage, leading the other two models by 22+ socres. As presented in Fig. 2 (1), in the leaderboard of the whole perception recognition, WeMM, InfMLLM, and SPHINX come in top three, closely followed by Lion, LLaVA, and XComposer-VL.

3.1.2 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” GPT-4V, WeMM, and XComposer-VL exceed other MLLMs, especially GPT-4V, which gets a score of 142.14. With respect to numerical calculation, GPT-4V still achieves first place, but falls short in the text translation. Regardless of whether it is commonsense reasoning, numerical calculation, or text translation, none of the highest scores exceed 150. This suggests that MLLMs have a lot of room for improvement in these capabilities. For the code reasoning, GPT-4V achieves a high score of 170, far ahead of other counterparts. For all of the cognition tasks, GPT-4V, Lion, and WeMM win the gold, silver, and bronze medals respectively, as shown in Fig. 2 (2).

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" [13], 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 and 2. 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. 30 advanced MLLMs are evaluated 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. Although MME is a comprehensive benchmark, it still has room for improvement in terms of capability coverage, such as covering more scenarios that require reasoning. We will continue to iterate MME series in the future to meet the evaluation requirements of MLLMs. More importantly, we hope that through the design of benchmarks, we can reflect the thoughts on the next capabilities of the model and thereby promote its development.

Acknowledgments

This work is funded by National Natural Science Foundation of China (Grant No. 62506158 and No. 62441234), Fundamental Research Funds for the Central Universities, AI & AI for Science Project of Nanjing University (No. 2024300529), and CCF-Tencent Rhino-Bird Open Research Fund.

References

[1] Infmllm. `https://github.com/mightyzau/InfMLLM`, 2023.

[2] Lion. `https://github.com/mynameischaos/Lion`, 2023.

[3] Octopus. `https://github.com/gray311/UnifiedMultimodalInstructionTuning`, 2023.

[4] Skywork-mm. `https://github.com/will-singularity/Skywork-MM`, 2023.

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

[6] Wemm. `https://github.com/scenarios/WeMM`, 2023.

[7] Xcomposer-vl. `https://github.com/InternLM/InternLM-XComposer`, 2023.

[8] 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.

 https://doi.org/10.52202/085713-4899

[9] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint:2308.12966*, 2023.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] Xinyu Fang, Kangrui Mao, Haodong Duan, Xiangyu Zhao, Yining Li, Dahua Lin, and Kai Chen. Mmbench-video: A long-form multi-shot benchmark for holistic video understanding. *NeurIPS*, 2024.

[15] 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.

[16] 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.

[17] 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.

[18] Jiaming Han, Renrui Zhang, Wenqi Shao, Peng Gao, Peng Xu, Han Xiao, Kaipeng Zhang, Chris Liu, Song Wen, Ziyu Guo, et al. Imagebind-llm: Multi-modality instruction tuning. *arXiv preprint:2309.03905*, 2023.

[19] Wenbo Hu, Yifan Xu, Y Li, W Li, Z Chen, and Z Tu. Bliva: A simple multimodal llm for better handling of text-rich visual questions. *arXiv preprint:2308.09936*, 2023.

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

[21] 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.

[22] 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.

[23] 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.

[24] Juncheng Li, Kaihang Pan, Zhiqi Ge, Minghe Gao, Hanwang Zhang, Wei Ji, Wenqiao Zhang, Tat-Seng Chua, Siliang Tang, and Yueting Zhuang. Fine-tuning multimodal llms to follow zero-shot demonstrative instructions. *arXiv preprint:2308.04152*, 2023.

[25] 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.

[26] 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.

[27] 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.

[28] Ziyi Lin, Chris Liu, Renrui Zhang, Peng Gao, Longtian Qiu, Han Xiao, Han Qiu, Chen Lin, Wenqi Shao, Keqin Chen, et al. Sphinx: The joint mixing of weights, tasks, and visual embeddings for multi-modal large language models. *arXiv preprint:2311.07575*, 2023.

[29] Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. Aligning large multi-modal model with robust instruction tuning. *arXiv preprint:2306.14565*, 2023.

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

[31] Jingying Liu, Binyuan Hui, Kun Li, Yunke Liu, Yu-Kun Lai, Yuxiang Zhang, Yebin Liu, and Jingyu Yang. Geometry-guided dense perspective network for speech-driven facial animation. *IEEE TVCG*, 2021.

[32] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? *arXiv preprint:2307.06281*, 2023.

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

[34] Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. *NeurIPS*, 2022.

[35] 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.

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

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

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

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

[40] 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.

[41] 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.

[42] Kimi Team, Angang Du, Bohong Yin, Bowei Xing, Bowen Qu, Bowen Wang, Cheng Chen, Chenlin Zhang, Chenzhuang Du, Chu Wei, et al. Kimi-vl technical report. *arXiv preprint arXiv:2504.07491*, 2025.

[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] Xiaoguang Tu, Zhi He, Yi Huang, Zhi-Hao Zhang, Ming Yang, and Jian Zhao. An overview of large ai models and their applications. *Visual Intelligence*, 2024.

[45] Jianfeng Wang, Zhengyuan Yang, Xiaowei Hu, Linjie Li, Kevin Lin, Zhe Gan, Zicheng Liu, Ce Liu, and Lijuan Wang. Git: A generative image-to-text transformer for vision and language. *arXiv preprint:2205.14100*, 2022.

[46] 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.

[47] 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.

[48] Hao Wen, Hongbo Kang, Jian Ma, Jing Huang, Yuanwang Yang, Haozhe Lin, Yu-Kun Lai, and Kun Li. Dycrowd: Towards dynamic crowd reconstruction from a large-scene video. *IEEE TPAMI*, 2025.

[49] 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.

[50] Yang Wu, Shilong Wang, Hao Yang, Tian Zheng, Hongbo Zhang, Yanyan Zhao, and Bing Qin. An early evaluation of gpt-4v (ision). *arXiv preprint:2310.16534*, 2023.

 https://doi.org/10.52202/085713-4899

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

[52] Qinghao Ye, Haiyang Xu, Jiabo Ye, Ming Yan, Haowei Liu, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration. *arXiv preprint:2311.04257*, 2023.

[53] Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. A survey on multimodal large language models. *arXiv preprint:2306.13549*, 2023.

[54] Shukang Yin, Chaoyou Fu, Sirui Zhao, Tong Xu, Hao Wang, Dianbo Sui, Yunhang Shen, Ke Li, Xing Sun, and Enhong Chen. Woodpecker: Hallucination correction for multimodal large language models. *arXiv preprint:2310.16045*, 2023.

[55] Kaining Ying, Fanqing Meng, Jin Wang, Zhiqian Li, Han Lin, Yue Yang, Hao Zhang, Wenbo Zhang, Yuqi Lin, Shuo Liu, et al. Mmt-bench: A comprehensive multimodal benchmark for evaluating large vision-language models towards multitask agi. *arXiv preprint arXiv:2404.16006*, 2024.

[56] Tianyu Yu, Jinyi Hu, Yuan Yao, Haoye Zhang, Yue Zhao, Chongyi Wang, Shan Wang, Yinxv Pan, Jiao Xue, Dahai Li, et al. Reformulating vision-language foundation models and datasets towards universal multimodal assistants. *arXiv preprint:2310.00653*, 2023.

[57] Yan Zeng, Hanbo Zhang, Jiani Zheng, Jiangnan Xia, Guoqiang Wei, Yang Wei, Yuchen Zhang, and Tao Kong. What matters in training a gpt4-style language model with multimodal inputs? *arXiv preprint:2307.02469*, 2023.

[58] Ao Zhang, Hao Fei, Yuan Yao, Wei Ji, Li Li, Zhiyuan Liu, and Tat-Seng Chua. Transfer visual prompt generator across llms. *arXiv preprint:2305.01278*, 2023.

[59] Jinsong Zhang, Xiongzheng Li, Hailong Jia, Jin Li, Zhuo Su, Guidong Wang, and Kun Li. Logavatar: Local gaussian splatting for human avatar modeling from monocular video. *CAD*, 2025.

[60] Jinsong Zhang, Minjie Zhu, Yuxiang Zhang, Zerong Zheng, Yebin Liu, and Kun Li. Speechact: Towards generating whole-body motion from speech. *IEEE TVCG*, 2025.

[61] Haozhe Zhao, Zefan Cai, Shuzheng Si, Xiaojian Ma, Kaikai An, Liang Chen, Zixuan Liu, Sheng Wang, Wenjuan Han, and Baobao Chang. Mmicl: Empowering vision-language model with multi-modal in-context learning. *arXiv preprint:2309.07915*, 2023.

[62] 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.

[63] 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.

[64] 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.

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

[66] 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.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Section 2.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Section 5.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

https://doi.org/10.52202/085713-4899

Justification: The benchmark is not applicable to this.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Section 3.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: The data and code have been released.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [NA]

   Justification: The benchmark is not applicable to this.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [NA]

   Justification: The benchmark is not applicable to this.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.

 https://doi.org/10.52202/085713-4899

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [NA]

   Justification: The benchmark is not applicable to this.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: This work meets the requirements.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: The benchmark is not applicable to this.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to

generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.

- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The benchmark is not applicable to this.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All relevant works have been cited.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

https://doi.org/10.52202/085713-4899

Answer: [NA]

Justification: The benchmark is not applicable to this.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: The benchmark is not applicable to this.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: The benchmark is not applicable to this.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: The benchmark is not applicable to this.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.

https://doi.org/10.52202/085713-4899