



MM-Vet: Evaluating Large Multimodal Models for Integrated Capabilities

Weihao Yu^{1*} Zhengyuan Yang^{2*} Linjie Li² Jianfeng Wang² Kevin Lin²
Zicheng Liu² Xinchao Wang^{1†} Lijuan Wang^{2†}

¹National University of Singapore ²Microsoft Azure AI

weihao.yu@nus.edu xinchao@nus.edu.sg

{zhengyang,lindsey.li,jianfw,keli,zliu,lijuanw}@microsoft.com

Abstract

We propose MM-Vet¹, an evaluation benchmark that examines large multimodal models (LMMs) on complicated multimodal tasks. Recent LMMs have shown various intriguing abilities, such as solving math problems written on the blackboard, reasoning about events and celebrities in news images, and explaining visual jokes. Rapid model advancements pose challenges to evaluation benchmark development. Problems include: (1) How to systematically structure and evaluate the complicated multimodal tasks; (2) How to design evaluation metrics that work well across question and answer types; and (3) How to give model insights beyond a simple performance ranking. To this end, we present MM-Vet, designed based on the insight that the intriguing ability to solve complicated tasks is often achieved by a generalist model being able to integrate different core vision-language (VL) capabilities. MM-Vet defines 6 core VL capabilities and examines the 16 integrations of interest derived from the capability combination. For evaluation metrics, we propose an LLM-based evaluator for open-ended outputs. The evaluator enables the evaluation across different question types and answer styles, resulting in a unified scoring metric. We evaluate representative LMMs on MM-Vet, providing insights into the capabilities of different LMM system paradigms and models. Code and data are available at <https://github.com/yuweihao/MM-Vet>.

1 Introduction

The breakthroughs in large language models (LLMs) [8, 54, 16, 3, 66, 32] bring generalist AI models that can solve a wide range of complicated natural language tasks, many approaching the human-expert-level performance [54, 10]. Large multimodal models (LMMs) aim to achieve even stronger general intelligence via extending LLMs with multimodal inputs. Since more than 80% of our human being’s perception, learning, cognition, and activities are mediated through vision [58], it is natural to start the exploration by equipping LLMs with “eyes.” One main thread of LMM works, represented by Frozen [68], Flamingo [2], PaLM-E [21], GPT-4 [54], extend LLMs with the visual understanding capability via end-to-end tuning. There also exists the exploration [75, 80, 76, 63, 26] on the modular combination of LLMs and image-to-text vision-language models. Recently, thanks to the open-source of powerful LLMs like LLaMA [66], more open-sourced LMMs are built, including OpenFlamingo [6], LLava [46], MiniGPT-4 [84], Otter [40], InstructBLIP [19], and many more [29, 47, 78]. These studies showcase the intriguing ability to solve various complicated

*Equal contribution.

†Corresponding authors.

¹Short for “Multimodal Veterinarian.”

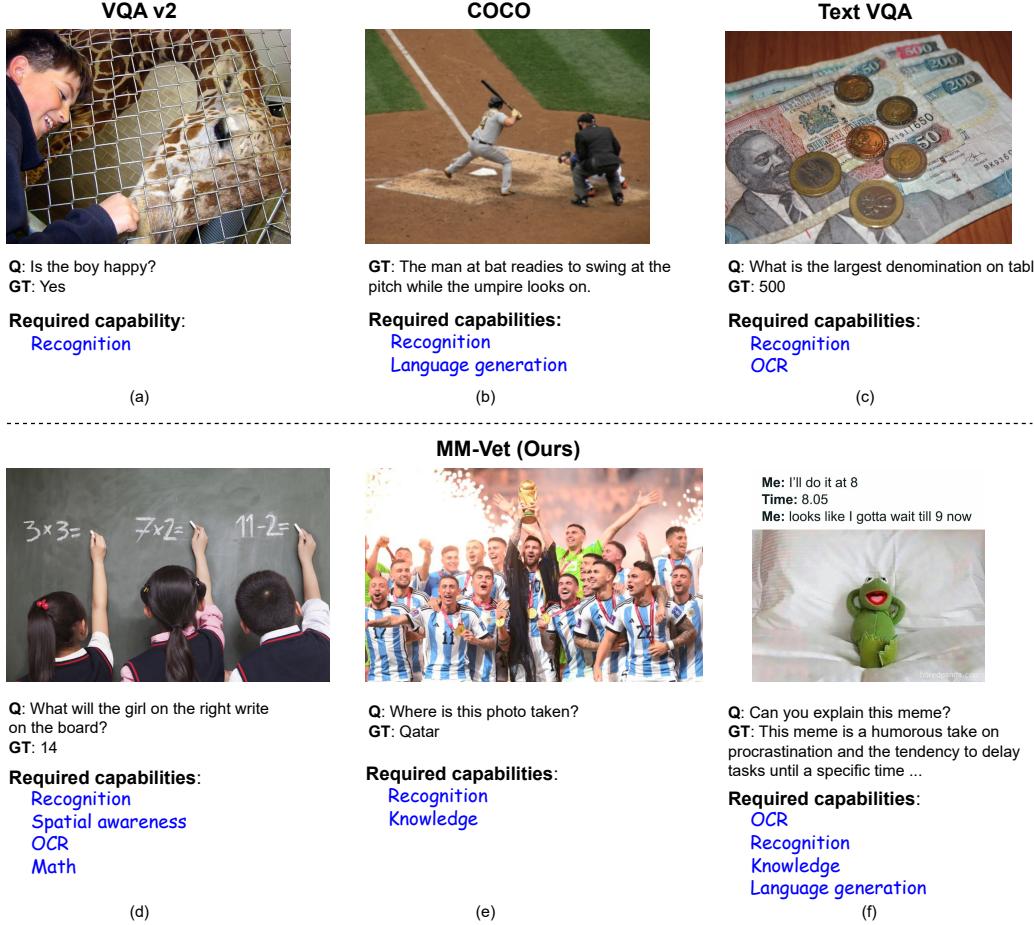


Figure 1: Required capabilities of different benchmarks. Different from conventional VL benchmarks only require one or two capabilities, MM-Vet focuses on the integration of different core VL capabilities, including recognition, OCR, knowledge, language generation, spatial awareness, and math.

multimodal tasks, such as open-world recognition, multimodal knowledge and commonsense, scene text understanding, and so on.

Despite the promising qualitative results on LMM’s capabilities, it remains unclear how to systematically evaluate those showcased complicated multimodal tasks and what are the relationships among evaluated tasks, which is the first step in developing a quantitative evaluation benchmark. As shown in Figure 1, existing vision-language benchmarks [4, 13, 65] focus on simple Vision-Language (VL) tasks that require specific one or two capabilities, such as recognition, language generation, or OCR, but fall short in benchmarking more complicated tasks. Alternatively, we examine the arbitrary integration of core VL capabilities for complicated tasks, with the insight that the intriguing ability to solve complicated multimodal tasks can be achieved by a generalist model mastering and integrating different core capabilities. Following this insight, we propose a new benchmark for evaluating LMMs, namely MM-Vet. MM-Vet defines six core VL capabilities, including recognition, OCR, knowledge, language generation, spatial awareness, and math, which integrate to solve various complicated multimodal tasks. MM-Vet contains 16 tasks for quantitative evaluation. For example, in Figure 1(d), answering the question “*What will the girl on the right write on the board?*” in MM-Vet requires recognizing the genders of the three kids, locating queried girl spatially, recognizing the scene text written by the girl, and finally calculating the result.

Other than the evaluation category definition, the evaluation metrics are another challenge in benchmark development, given the diverse answer styles and question types. Specifically: (1) The desired outputs in different multimodal tasks have diverse formats, e.g., Figure 1(d)’s math problem can be

answered by a single word, while outputs for the essay writing question are hundred-words long; **(2)** The core aspect to evaluate in different tasks varies, *e.g.*, text generation focuses more on the text quality, recognition can be considered correct with the key concept recognized. Most integrated tasks would require comprehensive evaluations from multiple dimensions. Inspired by recent NLP studies [15, 48, 24] that use LLMs for model evaluation, we propose an LLM-based evaluator as the evaluation metric for open-ended model outputs. As shown in Table 1, we prompt GPT-4 [54] with few-shot evaluation prompts to obtain an evaluation score ranging from 0 to 1. Instead of manually defining the possible answer styles and question types, we include different sample types as few-shot examples and let LLMs infer the scoring criteria automatically. Such metric design eases the future extension to more question types, such as box localization [12, 74, 70].

MM-Vet’s evaluation category and metric designs allow users to obtain capability insights for different LMMs. Such model analyses are more informative than a single overall ranking, which highly depends on the dataset sample composition and might be biased. We evaluate two sets of multimodal systems, *i.e.*, the end-to-end tuned LMMs including OpenFlamingo [6], LLaVA [46], MiniGPT-4 [84], Otter [40], InstructBLIP [19], *etc*, and the LLM-tool-using systems [76, 63, 26, 34] such as MM-ReAct [76]. Despite not knowing model details, we also evaluate industry solutions such as Bard [30]. We first discuss the capability analyses of these two system paradigms and the representative models. We then dive deeper into the open-sourced LMMs and examine how the training data, vision encoder, and LLM selection influence the performance on different capabilities.

Our contributions are summarized as follows.

- We propose MM-Vet to evaluate LMMs’ ability on complicated multimodal tasks. MM-Vet defines 16 emergent tasks of interest, integrated from the six defined core VL capabilities.
- We propose an LLM-based evaluator for open-ended outputs of LMMs, which unifies the evaluation across different answer styles and question types. The evaluation metrics ensure the thorough evaluation of both the factual correctness and text quality of the responses.
- We benchmark representative LMMs on MM-Vet, revealing the relative strengths and weaknesses of different system paradigms and models, as summarized in Section 4.5.

2 Related work

Multimodal models. Vision-language models [13, 31, 51, 14, 44, 36, 72, 69, 74, 25] approach multimodal intelligence of jointly understanding and generating vision and language signals. Inspired by the impressive quality and genericity in recent large language models (LLMs) [9, 54, 16, 66], researchers explore large multimodal models (LMMs) that seamlessly integrate different vision-language capabilities to solve complicated multimodal tasks. In approaching such multimodal generalist systems, one direction is to extend LLMs with the multi-sensory ability, such as pioneer works Frozen [68], Flamingo [2], PaLM-E [21], GPT-4 [54]. Recent open-sourced LLMs [82, 66, 57] also facilitate various research studies including OpenFlamingo [6], LLaVA [46], MiniGPT-4 [84], Otter [40], InstructBLIP [19], and so on [29, 47, 78]. On the other hand, multimodal agents [76, 63, 34, 26] explore chaining different vision tools with LLMs [9, 54] to achieve integrated vision-language capabilities.

VL benchmarks. Classic VL benchmarks focus on specific capabilities of interest, such as visual recognition [31], image description [13, 1], as well as other benchmarks for specialized capabilities such as scene text understanding [65, 64, 77], commonsense reasoning [79], outside knowledge [53]. The recent development of generalist LMMs posts a strong need for modernized VL benchmarks, which contain complicated multimodal tasks that require integrated VL capabilities.

Our MM-Vet is most related to the concurrent evaluation studies [23, 49, 41, 73] such as MME and MMBench, which design comprehensive evaluation samples to facilitate the LMM evaluation. One major difference is that MM-Vet defines and studies the integrated VL capabilities, allowing the evaluation to provide insights beyond the overall model ranking.

LLM-based evaluation. MM-Vet adopts the open-ended LLM-based evaluator, allowing the evaluation across answer styles and question types without requiring binary or multiple answer choices. The technique of prompting LLMs for model evaluation is related to the explorations in NLP [15, 48, 24]. We show that the technique extends well to multimodal tasks, and presents a unified prompt to evaluate samples with different answer styles and question types.

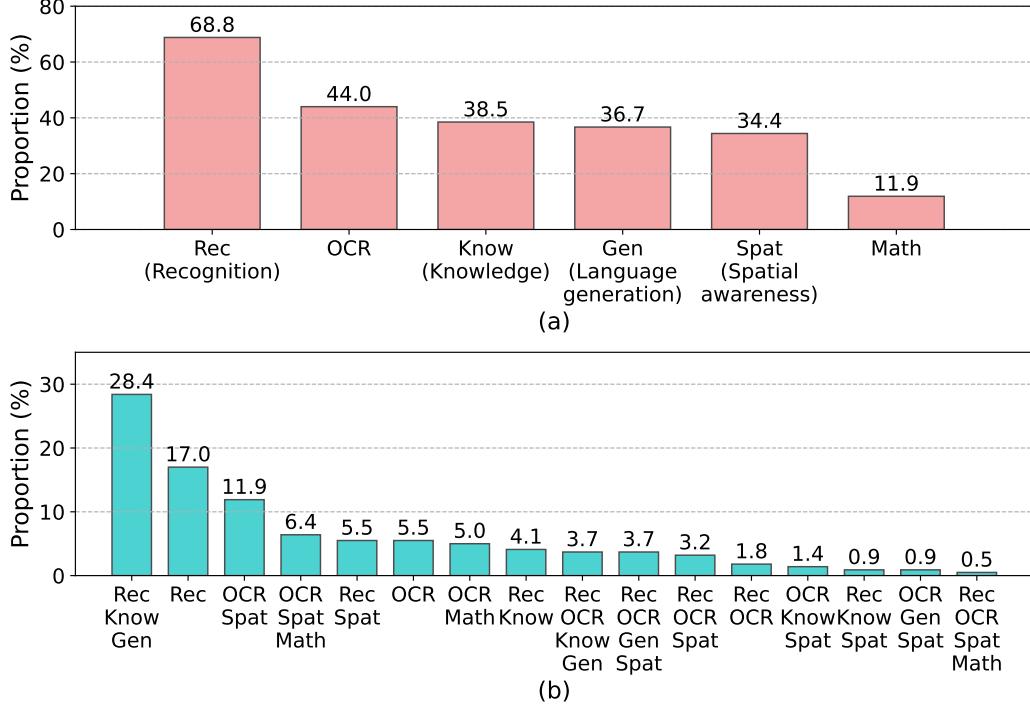


Figure 2: MM-Vet proportion of capabilities. (a) The proportion of each capability. The sum of the proportion is larger than 100% because most samples have more than one capability. (b) The proportion of capability integrations. The sum of the proportion is equivalent to 100%.

3 MM-Vet

3.1 Data collection

Our aim is to develop a multimodal benchmark that requires comprehensive capabilities, corresponding to realistic scenarios an AI agent might encounter. Consider, for instance, this scenario: Awakening from slumber, you reach out for your smartphone (*recognition capability*) to check the current time (*OCR capability*). Today, your plan is to visit a new grocery that you have not been to. Guided by the information that the grocery is situated directly opposite the stadium and next to the cinema (*spatial awareness*), you manage to locate it successfully. Keeping in mind your doctor’s advice to shed some weight, you consciously steer clear of high-calorie food and choose milk, vegetables, and fruits instead (*knowledge capability*). In the dairy aisle, you’re faced with a choice between two types of pure milk. The first is 4 dollars for one liter with 20% discount, while the second is 7 dollars for 1.5 liter with 25% discount. After some quick arithmetic, you find the former is cheaper (*math capability*) and opt for the one-liter package. After shopping, you walk past the cinema and find a person pointing to the poster to introduce a new movie (*language generation*).

From the scenarios of interest, we summarize the following six core VL capabilities for evaluation, with corresponding MM-Vet examples shown in Tables 8-13.

- **Recognition (Rec).** Recognition refers to the general visual recognition capability, including recognizing scenes, objects, object attributes (color, material, shape, *etc*), counting, and various other high-level visual recognition tasks in computer vision.
- **Knowledge (Know).** The knowledge category covers various knowledge-related capabilities, including social and visual commonsense knowledge, encyclopedic knowledge, and time-sensitive knowledge like news. This capability necessitates that the model not only possesses such knowledge, but also effectively utilizes it to solve complicated tasks as required.
- **OCR.** Optical character recognition (OCR) refers to the scene text understanding and reasoning capability. The models are tested to read the scene text in images, and reason over the texts to solve various tasks.

- **Spatial awareness (Spat).** Spatial awareness embodies a diverse spectrum of capabilities related to understanding space, including the comprehension of the spatial relationship among object and scene text regions.
- **Language generation (Gen).** Language generation is a vital ability that empowers models to articulate their responses in a clear, engaging, and informative manner. We use questions that demand more extended answers for language generation capacity evaluation.
- **Math.** Math evaluates the model’s arithmetic capability in solving either written equations or problems in the wild.

In real-world scenarios, various complicated multimodal tasks would require the integrations of different core VL capabilities. For instance, explaining visual jokes as shown in Table 8(a) requires recognition, knowledge of humor, and language generation; reading documents and solving math problems as shown in Table 9(a) takes OCR, spatial awareness and math; and answering exam questions given images as shown in Table 12(b) needs OCR, knowledge, spatial awareness. To solve these complicated tasks, LMMs are expected to seamlessly integrate different VL capabilities. Therefore, it is crucial to establish a benchmark that evaluates the performance of these integrated abilities within LMMs.

To build the benchmark, we have gathered 187 images from various online sources and ask 205 questions, each of which requires one or more capabilities to answer. As shown in Tables 8–13, these questions are varied in type and entail open-ended responses of differing lengths. The ground truths for 155 questions are human-annotated, while the remainder of the answers for 50 questions were gathered from the Internet. In addition to the 187 images, ten extra images with high-quality questions are collected from VCR [79], with the questions and answers modified to an open-ended answering format. Another three images are from ChestX-ray14 [71] to obtain corresponding medical expert knowledge. In total, our MM-Vet contains 200 images, and 218 questions (samples), all paired with their respective ground truths. For each question, we have also identified the capacities required to answer them and displayed this information statistically in Figure 2.

3.2 LLM-based evaluator for open-ended model outputs

Questions and expected responses in MM-Vet are designed to be open-ended to cover the diverse real-world scenarios. This naturally poses a great challenge in terms of model evaluation and metric design. Drawing inspiration from recent NLP studies [15, 83] that utilize LLMs for open-ended evaluations, we leverage GPT-4 to assist evaluation. As shown in Table 1, we craft a few-shot prompt for model evaluation. The few-shot design allows us to define the scoring metrics via in-context examples and supports easy extension onto new problem sets. Specifically, our implemented prompt incorporates five in-context examples with open-ended short answers and two examples with long answers. We cover examples that are fully correct (*i.e.*, 1.0) or incorrect (*i.e.*, 0.0), as well as examples used to define different types of “partially correct” responses. The LLM-based evaluator allows any style of model outputs to be evaluated with a unified consistent metric. Furthermore, it also supports easy adaptation to diverse question types and answer styles by simply modifying the evaluation examples.

By inputting the prompt, GPT-4 automatically generates scores for each sample, conditioned on each sample’s input question, ground truth, and model output. The score for each sample ranges from 0 to 1. The total scores are computed by

$$S = \frac{\sum_{i=1}^N s_i}{N} \times 100\%, \quad (1)$$

where s_i is the score of sample i , and N is the sample number. The score regarding each capability or capability integration can be similarly obtained by

$$S_c = \frac{\sum s_i}{N_c} \times 100\%, \quad i \in C, \quad (2)$$

where C is the set of samples requiring a specific capability or capability integration, and N_c is the sample number of the set.

Table 1: Few-shot prompt for evaluating model outputs using GPT-4, where \mathcal{Q} is a sample’s question, \mathcal{G} is the ground truth and \mathcal{P} is the model output for the sample. In the prompt, there are examples with short and long open-ended answers, enabling the evaluation of diverse answer styles. Taking the prompt filled with \mathcal{Q} , \mathcal{G} and \mathcal{P} , GPT-4 will generate a soft grading score from 0 to 1.

Compare the ground truth and prediction from AI models, to give a correctness score for the prediction. <AND> in the ground truth means it is totally right only when all elements in the ground truth are present in the prediction, and <OR> means it is totally right when any one element in the ground truth is present in the prediction. The correctness score is 0.0 (totally wrong), 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, or 1.0 (totally right). Just complete the last space of the correctness score.				
Question Ground truth Prediction Correctness				
— — — —				
What is x in the equation? -1 <AND> -5 x = 3 0.0				
What is x in the equation? -1 <AND> -5 x = -1 0.5				
What is x in the equation? -1 <AND> -5 x = -5 0.5				
What is x in the equation? -1 <AND> -5 x = -5 or 5 0.5				
What is x in the equation? -1 <AND> -5 x = -1 or x = -5 1.0				
Can you explain this meme? This meme is poking fun at the fact that the names of the countries Iceland and Greenland are misleading. Despite its name, Iceland is known for its beautiful green landscapes, while Greenland is mostly covered in ice and snow. The meme is saying that the person has trust issues because the names of these countries do not accurately represent their landscapes. The meme talks about Iceland and Greenland. It’s pointing out that despite their names, Iceland is not very icy and Greenland isn’t very green. 0.4				
Can you explain this meme? This meme is poking fun at the fact that the names of the countries Iceland and Greenland are misleading. Despite its name, Iceland is known for its beautiful green landscapes, while Greenland is mostly covered in ice and snow. The meme is saying that the person has trust issues because the names of these countries do not accurately represent their landscapes. The meme is using humor to point out the misleading nature of Iceland’s and Greenland’s names. Iceland, despite its name, has lush green landscapes while Greenland is mostly covered in ice and snow. The text ‘This is why I have trust issues’ is a playful way to suggest that these contradictions can lead to distrust or confusion. The humor in this meme is derived from the unexpected contrast between the names of the countries and their actual physical characteristics. 1.0				
$\mathcal{Q} \mathcal{G} \mathcal{P} $				

4 Evaluation results

4.1 Experiment settings

We utilize MM-Vet to evaluate two types of LMMs, *i.e.*, (1) end-to-end tuned LMMs (OpenFlamingo [2, 6], BLIP-2 [42], LLaVA [46], MiniGPT-4 [84], Otter [40] and InstructBLIP [19]); (2) LLM-tool-using methods (MM-ReAct [76] and Transformers Agent [34]). The summary of these methods is shown in Table 2. As shown in Table 1, for each sample, we fill the prompt template with its question, ground truth, and output from a specific LMM. By taking the filled prompt into GPT-4, GPT-4 will generate a score from 0 to 1 for the sample. It is found that outputs of GPT-4 still exist variance although the temperature is set as 0. Therefore, we utilize GPT-4 to evaluate the outputs of LLMs by 5 times. Due to the space limit, we report average scores for capabilities/capability integrations, and average as well as variance for total score.

4.2 Result analyses

The main results of different methods are shown in Table 3 regarding each capability, and Table 4 for each capability integration.

Table 2: Summary of the evaluated LMMs in this report. We consider both the end-to-end tuned models (*i.e.*, OpenFlamingo [2, 6], BLIP-2 [42], LLaVA [46], MiniGPT-4 [84], LLaMA-Adapter v2 [28], Otter [40], InstructBLIP [19]), and the LLM-tool-using systems (*i.e.*, MM-ReAct [76] and Transformers Agent [34]).

Method	Initial models			Tuning data	Total params
	Vision	Language	Other		
OpenFlamingo-9B [2, 6]	CLIP ViT-L/14 [59]	LLaMA-7B [66]	–	Multimodal C4 [85]	9B
BLIP-2-12B [42]	EVA-ViT-G [22]	Flan-T5-XXL [17]	–	1. COCO [45]; 2. Visual Genome [37]; 3. CC3M [62]; 4. CC12M [11]; 5. SBUs [55]; 6. 115M images from the LAION-400M [61]. (CapFilt [43] is used to create synthetic captions for the web images)	12B
LLaVA-7B [46]	CLIP ViT-L/14 [59]	Vicuna-7B [83]	–	1. CC3M [62] Concept-balanced 595K [46]; 2. LLaVA-Instruct-158K [46].	7B
LLaVA-13B [46]		Vicuna-13B [83]	–		13B
LLaVA-7B (LLaMA-2) [46]	CLIP ViT-L/14 [59]	LLaMA-2-7B-Chat [67]	–	1. LAION /CC/SBU BLIP-Caption Concept-balanced 558K [46];	7B
LLaVA-13B (LLaMA-2) [46]	CLIP ViT-L/14 [59]	LLaMA-2-13B-Chat [67]	–	2. LLaVA-Instruct-80K [46].	13B
LLaVA-13B (V1.3, 336px) [46]	CLIP ViT-L/336px [59]	Vicuna-13B-v1.3 [83]	–		13B
MiniGPT-4-8B [84]	EVA-ViT-G [22]	Vicuna-7B [83]	BLIP-2’s Q-Former [42]	1. CC3M [62]; 2. CC12M [11]; 3. SBUs [55]; 4. LAION-400M [61] 5. Proposed 3,500 aligned image-text pairs [84].	8B
MiniGPT-4-14B [84]		Vicuna-13B [83]	BLIP-2’s Q-Former [42]		14B
LLaMA-Adapter v2-7B [28]	CLIP ViT-L/14 [59]	LLaMA-7B [66]	–	1. GPT-4-LLM [57]; 2. COCO [45]	7B
Otter-9B [40]	CLIP ViT-L/14 [59]	LLaMA-7B [66]	1. Perceiver Resampler; 2. GATED XATTN-DENSE	MIMIC-IT [39]	9B
InstructBLIP-8B [19]		Vicuna-7B [83]	BLIP-2’s Q-Former [42]	1. Tuning data of BLIP-2 [42]; 2. 26 publicly available datasets (transformed into instruction tuning format).	8B
InstructBLIP-14B [19]		Vicuna-13B [83]	BLIP-2’s Q-Former [42]		14B
Transformers Agent (GPT-4 as agent) [34]	–	1. GPT-4 [54]; 2. Flan-T5 [17]; 3. BART [38]	1. Donut [35]; 2. BLIP [43]; 3. ViLT [36]; 4. CLIPSeg [52] 5. Whisper [60]; 6. SpeechT5 [5]; 7. NLLB [18]	None	Not clear
MM-ReAct-GPT-3.5 [76]	–	GPT-3.5 [56]	1. Azure Cognitive Services APIs [7] for image captioning, image tagging, dense captioning, OCR and specialized recognition on celebrities, receipts, <i>etc</i>	None	Not clear
MM-ReAct-GPT-4 [76]	–	GPT-4 [54]	2. Bing search; 3. PAL [27]		

4.2.1 Regarding each capability

Recognition. The “Recognition” category contains the questions requiring recognition capability to answer. Examples are shown in Tables 8(a, b), 9(b), 10(a, b), 11(a, b), 12(a, c), and 13(b). The “Rec” column in Table 3 compares the performance on the “Recognition”. Among the evaluated models, LLaVA-13B (LLaMA-2) is the best one, obtaining 39.2%. There may be two reasons. First, LLaVA-13B (LLaMA-2) adopts ViT-L/14 [20] from CLIP [59] as a vision model, which is trained by a large amount of data, 400 million image-text pairs; 2) Second, it is surprising that stronger language model can largely boost the recognition performance. LLaVA-13B (LLaMA-2) obtains 8.3% important over LLaVA-13B (Vicuna-13B). Stronger LLMs may help understand questions better and identify key information from visual inputs.

Besides, for model parameters below 10B, InstructBLIP-8B [19] attains the best performance (32.4% in MM-Vet). As shown in Table 2, the tuning data of InstructBLIP includes 26 publicly available datasets, which contain recognition heavily datasets, like VQA v2 [31] and GQA [33]. The promising capability of InstructBLIP in recognition may benefit from these datasets.

OCR. OCR assesses models’ capabilities in recognizing scene texts in images and performing various types of reasoning including math, spatial, recognition, *etc*. Examples are shown in Tables 8(c), 9(a, c, d), 10(b), 11(a, b), 12(a, b), 13(a, b). As shown in Table 2’s “OCR” column, MMReAct-GPT4 [76] performs the best (65.7%) in OCR capability with the assistance of an external OCR model as a tool. Among end-to-end tuned models, LLaVA-13B (LLaMA-2) [46] achieves the highest performance (22.7%). This superior performance may be attributed to LLaVA’s adoption of CLIP [59] ViT-L/14 [20] as its vision model, and the inclusion of a large volume of image-OCR pairings within the training data [50].

Knowledge. As depicted in Tables 8(a), 10(a, b) and 12(b, c), the “knowledge” category covers a wide range of knowledge-related questions, ranging from joke understanding to encyclopedia knowledge. MMReAct-GPT4 [76] achieves the best score in this capability as shown in Table 3, because of its strong LLM backbone [54], coupled with external tools like Bing search for knowledge acquisition.

Table 3: MM-Vet evaluation results on various LMMs regarding each *core VL capability*. For each column, the highest, the second, and the third highest figures are highlighted by green, orange and blue backgrounds. All the numbers are presented in % and the full score is 100%.

Model	Rec	OCR	Know	Gen	Spat	Math	Total
Transformers Agent (GPT-4) [34]	18.2	3.9	2.2	3.2	12.4	4.0	13.4±0.5
LLaMA-Adapter v2-7B [28]	16.8	7.8	2.5	3.0	16.6	4.4	13.6±0.2
OpenFlamingo-9B [2, 6]	24.6	14.4	13.0	12.3	18.0	15.0	21.8±0.1
MiniGPT-4-8B [84]	27.4	15.0	12.8	13.9	20.3	7.7	22.1±0.1
BLIP-2-12B [42]	27.5	11.1	11.8	7.0	16.2	5.8	22.4±0.2
LLaVA-7B [46]	28.0	17.1	16.3	18.9	21.2	11.5	23.8±0.6
MiniGPT-4-14B [84]	29.9	16.1	20.4	22.1	22.2	3.8	24.4±0.4
Otter-9B [40]	28.4	16.4	19.4	20.7	19.3	15.0	24.6±0.2
InstructBLIP-14B [19]	30.8	16.0	9.8	9.0	21.1	10.5	25.6±0.3
InstructBLIP-8B [19]	32.4	14.6	16.5	18.2	18.6	7.7	26.2±0.2
LLaVA-13B [46]	30.9	20.1	23.5	26.4	24.3	7.7	26.4±0.1
MM-ReAct-GPT-3.5 [76]	24.2	31.5	21.5	20.7	32.3	26.2	27.9±0.1
LLaVA-7B (LLaMA-2) [46]	32.9	20.1	19.0	20.1	25.7	5.2	28.1±0.4
LLaVA-13B (V1.3, 336px) [46]	38.1	22.3	25.2	25.8	31.3	11.2	32.5±0.1
LLaVA-13B (LLaMA-2) [46]	39.2	22.7	26.5	29.3	29.6	7.7	32.9±0.1
MM-ReAct-GPT-4 [76]	33.1	65.7	29.0	35.0	56.8	69.2	44.6±0.2

Language generation. ‘‘Language generation’’ denotes the proficiency to produce fluent and informative text outputs, as illustrated in Table 8(a), 10(b), 11(a), and 13(a). The performance within this category is highly correlated with the efficacy of language modeling. As a result, MMReAct-GPT4 [76] and LLaVA-13B (LLaMA-2) stand out as the top two models. Their success can be attributed to the GPT-4 and LlaMA-2 language models on which these systems are built.

Spatial awareness. ‘‘Spatial awareness’’ involves the understanding of the spatial relationship among visual object regions (*e.g.*, Table 8(c)) and scene text regions (*e.g.*, Table 11(a, b)). MMReAct-GPT4 [76] has a significant lead in this capability (56.8%), because the adopted tools, such as dense captioning and OCR, provide detailed object and scene text location information in the form of coordinates, which can be understood and processed by GPT-4.

When it comes to end-to-end tuned models, LLaVA-13B (V1.3, 336px) exhibits the best performance of 31.3%. The tuning data for LLaVA is partly derived from capturing object names and their corresponding coordinates as input. This procedure ensures the generation of data imbued with spatial information, potentially aiding the models in developing and enhancing their spatial awareness capabilities.

Math. ‘‘Math’’ measures the arithmetic capability on either written equations (*e.g.*, Table 13(b)) or problems in the wild (*e.g.*, Table 9(d)). Notably, MMReAct-GPT4 [76] consistently outperforms other models. This superior performance may be attributed to the adopted PAL math tool (Program-aided Language Models) [27].

4.2.2 Regarding each capability integration

Recognition, knowledge, and language generation.. As shown in Table 8(a), this capability integration can enable models to explain visual jokes. LLaVA-13B (LLaMA-2) and LLaVA-13B (V1.3, 336px) [46] are the best models in this capability integration. Adopting CLIP [59] and stronger language models may be the reason. The tuning data of LLaVA shown in Table 2 can also not be ignored.

Recognition (sole). This category contains samples only requiring recognition, as shown in Table 8(b). InstructBLIP-14B and InstructBLIP-8B [19] achieve the best performance, which may result from the tuning data including recognition datasets, like VQA v2 [31] and GQA [33].

OCR and spatial awareness. For this integration, an example is shown in Table 8(c). MM-ReAct-GPT-4 [76] is the best method for this integration. Notably, compared with MM-ReAct-GPT-3.5,

Table 4: MM-Vet evaluation results on various LMMs regarding each *capability integration*. Examples of each capability integration are shown in supplementary materials Tables 8-13. For each column, the highest, the second, and the third highest figures are highlighted by green, orange and blue backgrounds. All the numbers are presented in % and the full score is 100%.

Model	Rec				OCR				Rec				OCR				Rec				Total	
	Know		Gen		OCR		Spat		Rec		OCR		Know		Gen		OCR		Know		Total	
	Know	Rec	Gen	Rec	OCR	Spat	Math	Rec	OCR	Math	Rec	Know	Gen	Spat	OCR	Know	Spat	Know	Gen	Spat	Math	
Transformers Agent (GPT-4) [34]	1.3	49.1	0.0	7.4	45.8	0.0	0.0	0.0	0.0	9.5	0.0	25.0	0.0	50.0	49.0	0.0	0.0	13.4±0.5				
LLaMA-Adapter v2-7B [28]	0.2	43.2	7.9	8.1	41.7	0.0	0.0	0.0	0.0	26.8	0.0	25.0	33.3	50.0	6.0	0.0	0.0	0.0	0.0	0.0	13.6±0.2	
OpenFlamingo-9B [2, 6]	15.6	48.6	17.3	21.4	41.7	18.3	8.2	11.1	2.5	0.0	14.3	50.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	21.8±0.1	
MiniGPT-4-8B [84]	14.2	47.9	9.6	14.3	50.0	20.8	0.0	14.4	8.0	21.2	42.9	50.0	0.7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	22.1±0.1	
BLIP-2-12B [42]	7.3	65.1	11.5	7.1	41.7	21.2	4.5	38.9	5.2	8.5	14.3	25.0	16.7	50.0	0.0	0.0	0.0	0.0	0.0	0.0	22.4±0.2	
LLaVA-7B [46]	17.1	46.6	13.3	21.4	41.7	24.8	0.0	28.9	6.2	45.2	6.6	50.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	23.8±0.6	
MiniGPT-4-14B [84]	21.1	47.5	14.6	7.1	50.0	16.7	0.0	11.1	18.7	38.5	18.3	32.5	50.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	24.4±0.4	
Otter-9B [40]	22.5	50.0	18.1	21.4	33.3	16.7	8.2	16.7	5.0	28.5	0.0	50.0	16.7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	24.6±0.2	
InstructBLIP-14B [19]	8.1	74.3	14.6	14.3	50.0	19.2	6.5	11.1	8.8	15.2	14.3	70.0	16.7	50.0	15.0	0.0	0.0	0.0	0.0	0.0	25.6±0.3	
InstructBLIP-8B [19]	18.0	69.9	15.4	14.3	33.3	20.8	0.0	23.3	7.8	35.2	15.7	25.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	26.2±0.2	
LLaVA-13B [46]	25.2	41.1	17.3	7.1	47.5	23.3	9.1	18.0	12.5	53.8	14.3	50.0	50.0	0.0	12.0	0.0	0.0	0.0	0.0	0.0	26.4±0.1	
MM-ReAct-GPT-3.5 [76]	19.1	33.1	28.8	35.7	28.3	60.0	9.1	33.3	2.5	47.8	0.0	25.0	100.0	0.0	0.0	35.0	80.0	0.0	0.0	0.0	27.9±0.1	
LLaVA-7B (LLaMA-2) [46]	18.8	57.0	26.9	9.7	50.0	26.7	0.0	34.7	10.2	44.8	14.3	50.0	11.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	28.1±0.4	
LLaVA-13B (V1.3, 336px) [46]	25.5	59.7	25.0	14.3	66.7	25.8	8.2	27.8	11.2	49.3	14.3	50.0	33.3	50.0	2.0	0.0	0.0	0.0	0.0	0.0	32.5±0.1	
LLaVA-13B (LLaMA-2) [46]	29.8	59.5	21.2	14.3	58.3	36.2	0.0	27.8	3.5	56.8	28.6	50.0	33.3	0.0	8.0	0.0	0.0	0.0	0.0	0.0	32.9±0.1	
MM-ReAct-GPT-4 [76]	22.5	33.0	69.2	78.6	25.0	83.0	63.6	44.4	68.2	88.0	14.3	50.0	0.0	50.0	80.0	0.0	0.0	44.6±0.2				

MM-ReAct-GPT-4 has a significant improvement, over 40%, indicating the importance of LLMs to integrate information of OCR and location.

OCR, spatial awareness, and math. An example of this integration is shown in Table 9(a), which requires reading the floor plan and conducting arithmetic. Compared with the above integration, this combination involves one more capability of math. The observation is similar to the integration of OCR and spatial awareness. MM-ReAct-GPT-4 [76] still achieves the best performance.

Recognition and spatial awareness. Table 9(b) shows an example for this integration. LLaVA-13B (V1.3, 336px) [46] performs best for this category. Compared with LLaVA-13B (LLaMA-2), LLaVA-13B (V1.3, 336px) obtains an improvement of 8.4%, indicating the significant contribution of larger resolution of images.

OCR (sole). This task requires OCR only, as shown in Table 9(c). MM-ReAct-GPT-4 [76] has the best results for sole OCR due to an OCR tool from Azure API. Notable, MM-ReAct-GPT-4 is much better than MM-ReAct-GPT-3.5 with an improvement of 23.0%, demonstrating the importance of language models in OCR.

OCR and Math. This integration enables reading text from real-world scenarios and solving math problems, as shown in Table 9(d). MM-ReAct-GPT-4 [76] obtains the best performance in this capability integration, far ahead of other models. We highly recommend using MM-ReAct-GPT-4 to complete tasks related to this capability integration.

Other capability integrations. 10 other capability integrations are in long-tailed distribution, where MMReAct-GPT-4 achieves the best scores in 6 integrations out of 10. Their examples are shown in Tables 10-13.

4.3 Result discussion

4.3.1 Foundation models and tuning data

In this subsection, we discuss the modules in LMMs and speculate how each component may affect the LMMs' capabilities in different aspects, evaluated by MM-Vet. We mainly consider the models based on open-sourced LLMs, *i.e.*, Flan-T5 [17], LLaMA [66], Vicuna [83], and LLaMA-2 [67].

Vision. For the Vision component, two models have been employed in the end-to-end LMMs we evaluated, namely, CLIP-ViT/L14 [59] (428M) and EVA-ViT-G (1.13B). Determining a superior model is currently not possible due to the absence of a comprehensive ablation study [81]. However, it's noteworthy that, when paired with the same language model, Vicuna-7B, InstructBLIP-8B excels in recognition tasks, while LLaVA-7B works particularly well for OCR.

Table 5: MM-Vet (Bard set) evaluation results on various LMMs regarding each *core VL capability*. For each column, the highest, the second, and the third highest figures are highlighted by green, orange and blue backgrounds. All the numbers are presented in % and the full score is 100%.

Model	Rec	OCR	Know	Gen	Spat	Math	Total
LLaVA-13B (LLaMA-2) [46]	37.8	22.9	22.4	27.6	27.2	8.0	30.3±0.1
LLaVA-13B (V1.3, 336px) [46]	39.4	22.3	22.7	24.6	30.6	11.6	31.5±0.1
MM-ReAct-GPT-3.5 [76]	22.3	31.4	15.6	16.6	32.9	24.0	27.6±0.2
MM-ReAct-GPT-4 [76]	34.3	66.3	25.6	36.6	60.6	72.0	48.1±0.2
Bard [30]	56.2	52.5	50.9	61.0	52.0	39.6	53.5±0.2

Table 6: MM-Vet (Bard set) evaluation results on various LMMs regarding each *capability integration*. For each column, the highest, the second, and the third highest figures are highlighted by green, orange and blue backgrounds. All the numbers are presented in % and the full score is 100%.

Model	Rec				OCR				Rec				OCR				Rec			
	Know		OCR		Spat		Rec		OCR		Rec		OCR		Rec		OCR		Rec	
	Gen	Rec	OCR	Spat	Math	Spat	Rec	OCR	Know	Gen	OCR	Rec	OCR	Know	Gen	Spat	Spat	Spat	Math	
Vicuna-13B (LLaMA-2) [46]	26.6	55.2	18.8	14.3	57.1	39.5	0.0	20.0	1.3	56.8	28.6	50.0	33.3	0.0	8.0	—	30.3±0.1			
Vicuna-13B (V1.3, 336px) [46]	21.9	59.0	22.9	14.3	85.7	25.5	8.2	20.0	15.0	49.3	14.3	50.0	33.3	50.0	2.0	—	31.5±0.1			
MM-ReAct-GPT-3.5 [76]	11.3	38.8	31.2	35.7	28.6	56.4	9.1	20.0	0.0	47.8	0.0	25.0	100.0	0.0	35.0	—	27.6±0.2			
MM-ReAct-GPT-4 [76]	17.0	35.2	70.8	78.6	28.6	81.5	63.6	40.0	68.3	88.0	14.3	50.0	0.0	50.0	80.0	—	48.1±0.2			
Bard [30]	52.3	70.3	45.2	56.4	42.9	70.2	18.2	0.0	77.7	81.5	28.6	50.0	66.7	50.0	80.0	—	53.5±0.2			

Language. There is a notable trend indicating that superior language models (LLMs) typically yield better performance, such as comparing the 7B and 13B variants of different models, except for the outlier of InstructBLIP where the 8B version performs better than the 14B one.

Tuning data. First and foremost, increasing the volume of data can significantly enhance performance. For example, Otter [40] is based on OpenFlamingo [6] with MIMIC-IT [39] to further tuning, and Otter gains clearly better results than OpenFlamingo. Another example is InstructBLIP-8B [19], which utilizes more data from 26 publicly available datasets to tune the model and achieve higher scores than BLIP-2-12B. Given the impressive performances of InstructBLIP and LLaVA, as demonstrated in Table 3, we expect further improvements by combining the tuning data of these two methods.

4.3.2 Comparison with Bard

Bard [30] is one popular closed-source commercial LMM system. One problem in evaluation is that Bard rejects images containing people and instead outputs “Sorry, I can’t help with images of people yet.” To conduct a fair comparison with other models, we constructed a subset of MM-Vet with 168 samples that Bard could process, henceforth referred to as the Bard set. The results on the Bard set are shown in Tables 5 and 6.

Bard achieves the highest scores in three out of six capabilities, seven out of fifteen capability integrations, and holds the highest overall score (53.5%). MM-ReAct-GPT-4 [76] outperforms in the remaining three out of six capabilities, and tops the chart in nine out of the fifteen capability integrations. Particularly, MM-ReAct performs better in OCR, spatial awareness, and math capabilities, indicating the potential benefit of having specialized external tools, even when working with state-of-the-art LMMs.

When considering end-to-end models, there is still a big gap from Bard. For instance, Vicuna-13B (V1.3, 336px) [46] obtains 31.5%, a substantial 22.0% lower than Bard. Future stronger open-sourced LLMs and advancements in multimodal training hold potential to further narrow this gap.

4.4 Effectiveness analysis of LLM-based evaluation

To verify the effectiveness of LLM-based evaluation for LMM predictions, we select the outputs from MMReAct-GPT-4 on 138 objective questions, which can be objectively annotated by humans. We compute the absolute value of the difference between the evaluator’s output score and the human-annotated score on each sample. The average of these absolute values is then computed to derive the final result, which is represented as $\bar{\Delta}$.

Table 7: Averaged absolute differences ($\bar{\Delta}$) between the evaluation scores of various LLM evaluators and those of human-annotated scores, on MM-ReAct-GPT4’s results. A smaller discrepancy indicates a better agreement with the gold standard of human evaluation, indicating a better evaluator.

Model	Keyword matching	LLM-based evaluation			
		LLaMA-2-7B	LLaMA-2-13B	GPT-3.5 (turbo-0613)	GPT-4 (0613)
$\bar{\Delta}$ (↓)	0.273	0.307	0.254	0.178	0.042

The maximum potential discrepancy is 1.0. The baseline evaluation method, keyword matching, results in a high difference of 0.273. This illustrates the unsuitability of keyword matching for MM-Vet when dealing with open-ended answers. It is surprising that $\bar{\Delta}$ of LLaMA-2-7B [67] is even higher than that of keyword matching, while $\bar{\Delta}$ LLaMA-2-13B only marginally less than keyword matching. This suggests that assessing open-ended outputs from models is far from straightforward. For OpenAI’s models, GPT-3.5 (turbo-0613) obtains 0.178 of $\bar{\Delta}$, and GPT-4 (0613) achieves the lowest difference of 0.042. In this paper, we utilize GPT-4 (0613) to evaluate the outputs of LMMs.

4.5 Takeaway notes

We summarize the above analyses and discussions as follows:

- In the evaluation of integrated capabilities on MM-Vet (Sections 4.2, 4.3.2), Bard [30] outperforms existing open-sourced methods. The tool-using approach, MM-ReAct-GPT-4 [76], achieves comparable performance to Bard with effective external tools. The pros and cons in different categories motivate future studies on tool-enhanced LMMs. Among end-to-end LMMs, LLaVA-13B (LLaMA-2)/LLaVA-13B (V1.3, 336px) [46] demonstrates the best performance on MM-Vet.
- Analysis of open-source LMMs (Section 4.3.1) leaves room for ambiguity regarding the superior vision encoders for LMMs, based on current model comparisons. However, it is evident that stronger LLMs can boost the performance of LMMs.
- For open-ended evaluation (Section 4.4), it is effective to use GPT-4 for evaluating the open-ended outputs of LMMs. The use of less powerful LLMs could result in more significant deviations from the gold standard of human evaluation results.
- Current top-performing methods, such as Bard [30] and MM-ReAct-GPT-4 [76], only achieve scores of around 50% on MM-Vet (where full score is 100%). The gap signifies that further effort is necessary to enhance the performance of LMMs in terms of integrated capabilities, *e.g.*, by developing stronger LMMs or extending LMMs with external tools.

5 Conclusion

In this paper, we introduce the MM-Vet benchmark to evaluate LMMs in terms of their integrated vision-language capabilities. We have assembled a new multimodal dataset, which requires the integration of multiple vision-language capabilities. To facilitate open-ended evaluation, we adopt an LLM-based evaluator to grade open-ended outputs from LMMs. We then evaluate various LMMs on MM-Vet, analyzing their results to provide insights into different LMM system paradigms and module selections. We observe that the current best LMMs only achieve around 50% scores on MM-Vet (full score 100%), indicating the need for efforts to further improve the integrated capabilities of LMMs.

A Examples of capability integrations

Table 8: Three samples requiring different capability integrations.

<p>(a)</p> 
<p>Q: What occasions would someone use this meme?</p> <p>GT: This meme, commonly known as "Screaming Panda," is typically used to express shock, surprise, or fear. It could be used in response to a startling or unexpected event, or to convey a sense of panic or alarm. Some possible occasions where someone might use this meme include:</p> <ul style="list-style-type: none"> - Reacting to a jump scare in a horror movie - Responding to a surprising plot twist in a TV show or book - Expressing shock at a news headline or current event - Conveying fear or anxiety about an upcoming deadline or exam - Showing surprise at an unexpected outcome in a sports game or other competition.
<p>Required capabilities: Recognition, knowledge, language generation</p>
<p>(b)</p> 
<p>Q: How many tomatoes are there?</p> <p>GT: 5</p>
<p>Required capabilities: Recognition</p>
<p>(c)</p> 
<p>Q: What is located to the right of the shampoo?</p> <p>GT: conditioner</p>
<p>Required capabilities: OCR, spatial awareness</p>

Table 9: Four samples requiring different capability integrations.

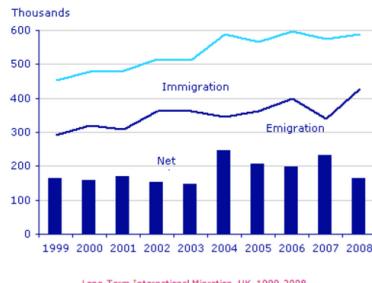
(a) 
<p>Q: Which room is bigger, the double garage or the living room?</p> <p>GT: double garage</p> <p>Required capabilities: OCR, spatial awareness, math</p>
(b) 
<p>Q: On the right desk, what is to the left of the laptop?</p> <p>GT: table lamp <OR> desk lamp</p> <p>Required capabilities: Recognition, spatial awareness</p>
(c) 
<p>Q: What are all the scene text in the image?</p> <p>GT: 5:30PM <AND> 88% <AND> Mario Kart 8 Deluxe <AND> MARIO KART 8 DELUXE <AND> SUPER MARIO ODYSSEY <AND> THE LEGEND OF ZELDA <AND> BREATH OF WILD <AND> Options <AND> Start</p> <p>Required capabilities: OCR</p>
(d) 
<p>Q: How many gallons of supreme gasoline can I get with \$50?</p> <p>GT: 13.6 <OR> 13.7</p> <p>Required capabilities: OCR, math</p>

Table 10: Two samples requiring different capability integrations.

(a)  <p>Q: In which country was this photo taken? GT: Australia Required capabilities: Recognition, knowledge</p>
(b) <p style="text-align: center;">Studying online</p> <div style="display: flex; justify-content: space-around; align-items: flex-start;"> <div style="text-align: center;">  <p>Zoom</p> </div> <div style="text-align: center;">  <p>Teacher</p> </div> </div> <div style="display: flex; justify-content: space-around; margin-top: 10px;"> <div style="text-align: center;">  <p>Wifi company</p> </div> <div style="text-align: center;">  <p>Students</p> </div> </div> <p>Q: Can you explain this meme? GT: This meme is a commentary on the current situation of online learning due to the COVID-19 pandemic. It shows how different parties are affected by it.</p> <p>The first two images show the Zoom company and the teachers throwing money in the air, which implies that they are profiting from the situation. Zoom, in particular, has seen a surge in usage and revenue due to the pandemic.</p> <p>The third image shows the WiFi company also throwing money in the air, which implies that they are also profiting from the situation as more people are using the internet for online learning.</p> <p>The last image shows a student with a sad clown face, which implies that students are not benefiting from the situation and are struggling with online learning. The "Put on a happy face" sign in the background further emphasizes the idea that students are putting on a brave face despite their struggles.</p> <p>Required capabilities: Recognition, OCR, knowledge, language generation</p>

Table 11: Two samples requiring different capability integrations.

(a)



Long-Term International Migration, UK, 1999-2008

- Q:** The graph below shows the long-term international migration, UK, 1999-2008. Summarize the information by selecting and reporting the main features, and make comparisons where relevant.
You should write at least 150 words.
- GT:** The chart gives information about UK immigration, emigration and net migration between 1999 and 2008.

Both immigration and emigration rates rose over the period shown, but the figures for immigration were significantly higher. Net migration peaked in 2004 and 2007.

In 1999, over 450,000 people came to live in the UK, while the number of people who emigrated stood at just under 300,000. The figure for net migration was around 160,000, and it remained at a similar level until 2003. From 1999 to 2004, the immigration rate rose by nearly 150,000 people, but there was a much smaller rise in emigration. Net migration peaked at almost 250,000 people in 2004.

After 2004, the rate of immigration remained high, but the number of people emigrating fluctuated. Emigration fell suddenly in 2007, before peaking at about 420,000 people in 2008. As a result, the net migration figure rose to around 240,000 in 2007, but fell back to around 160,000 in 2008.

Required capabilities: Recognition, OCR, language generation, spatial awareness

(b)



- Q:** Which car is on the parking spot 33?
GT: no <OR> empty

Required capabilities: Recognition, OCR, spatial awareness

Table 12: Three samples requiring different capability integrations.

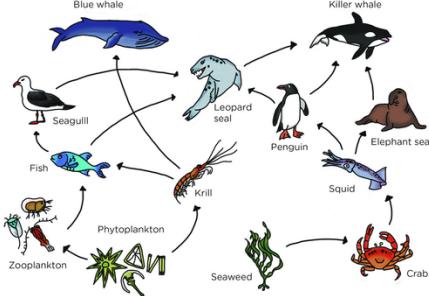
(a) 
Q: Is this apple organic? GT: yes Required capabilities: Recognition, OCR
(b) 
Q: Which are producers in this food web? GT: Phytoplankton <AND> Seaweed Required capabilities: OCR, knowledge, spatial awareness
(c) 
Q: Does the person bigger than the car? GT: no Required capabilities: Recognition, knowledge, spatial awareness

Table 13: Two samples requiring different capability integrations.

(a)

Underground Railway Systems

City	Date opened	Kilometres of route	Passengers per year (in millions)
London	1863	394	775
Paris	1900	199	1191
Tokyo	1927	155	1927
Washington DC	1976	126	144
Kyoto	1981	11	45
Los Angeles	2001	28	50

Q: The table below gives information about the underground railway systems in six cities.

Summarise the information by selecting and reporting the main features, and make comparisons where relevant.

You should write at least 150 words.

GT: The table shows data about the underground rail networks in six major cities.

The table compares the six networks in terms of their age, size and the number of people who use them each year. It is clear that the three oldest underground systems are larger and serve significantly more passengers than the newer systems.

The London underground is the oldest system, having opened in 1863. It is also the largest system, with 394 kilometres of route. The second largest system, in Paris, is only about half the size of the London underground, with 199 kilometres of route. However, it serves more people per year. While only third in terms of size, the Tokyo system is easily the most used, with 1927 million passengers per year.

Of the three newer networks, the Washington DC underground is the most extensive, with 126 kilometres of route, compared to only 11 kilometres and 28 kilometres for the Kyoto and Los Angeles systems. The Los Angeles network is the newest, having opened in 2001, while the Kyoto network is the smallest and serves only 45 million passengers per year.

Required capabilities: OCR, language generation, spatial awareness

(b)



Q: What will the girl on the right write on the board?

GT: 14

Required capabilities: Recognition, OCR, spatial awareness, math

References

- [1] Harsh Agrawal, Karan Desai, Yufei Wang, Xinlei Chen, Rishabh Jain, Mark Johnson, Dhruv Batra, Devi Parikh, Stefan Lee, and Peter Anderson. Nocaps: Novel object captioning at scale. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8948–8957, 2019.
- [2] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736, 2022.
- [3] Rohan Anil, Andrew M Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, et al. Palm 2 technical report. *arXiv preprint arXiv:2305.10403*, 2023.
- [4] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. VQA: Visual Question Answering. In *ICCV*, 2015.
- [5] Junyi Ao, Rui Wang, Long Zhou, Chengyi Wang, Shuo Ren, Yu Wu, Shujie Liu, Tom Ko, Qing Li, Yu Zhang, et al. Speecht5: Unified-modal encoder-decoder pre-training for spoken language processing. *arXiv preprint arXiv:2110.07205*, 2021.
- [6] 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, March 2023.
- [7] Microsoft Azure. Azure cognitive services apis. <https://azure.microsoft.com/en-us/products/ai-services/ai-vision>, 2023. Accessed: 2023-06-20.
- [8] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.
- [9] Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. In *NeurIPS*, 2020.
- [10] Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*, 2023.
- [11] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3558–3568, 2021.
- [12] Ting Chen, Saurabh Saxena, Lala Li, David J Fleet, and Geoffrey Hinton. Pix2seq: A language modeling framework for object detection. In *ICLR*, 2022.
- [13] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015.
- [14] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Learning universal image-text representations. In *ECCV*, 2020.
- [15] Cheng-Han Chiang and Hung-yi Lee. Can large language models be an alternative to human evaluations? *arXiv preprint arXiv:2305.01937*, 2023.
- [16] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*, 2022.
- [17] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*, 2022.
- [18] Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heffernan, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*, 2022.

- [19] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *arXiv preprint arXiv:2305.06500*, 2023.
- [20] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.
- [21] Danny Driess, Fei Xia, Mehdi S. M. Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, Wenlong Huang, Yevgen Chebotar, Pierre Serbanet, Daniel Duckworth, Sergey Levine, Vincent Vanhoucke, Karol Hausman, Marc Toussaint, Klaus Greff, Andy Zeng, Igor Mordatch, and Pete Florence. Palm-e: An embodied multimodal language model. In *arXiv preprint arXiv:2303.03378*, 2023.
- [22] Yuxin Fang, Wen Wang, Binhui Xie, Quan Sun, Ledell Wu, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. Eva: Exploring the limits of masked visual representation learning at scale. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19358–19369, 2023.
- [23] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Zhenyu Qiu, Wei Lin, Jinrui Yang, Xiawu Zheng, et al. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*, 2023.
- [24] Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. Gptscore: Evaluate as you desire. *arXiv preprint arXiv:2302.04166*, 2023.
- [25] Zhe Gan, Linjie Li, Chunyuan Li, Lijuan Wang, Zicheng Liu, and Jianfeng Gao. Vision-language pre-training: Basics, recent advances, and future trends. *arXiv preprint arXiv:2210.09263*, 2022.
- [26] Difei Gao, Lei Ji, Luowei Zhou, Kevin Qinghong Lin, Joya Chen, Zihan Fan, and Mike Zheng Shou. Assistgpt: A general multi-modal assistant that can plan, execute, inspect, and learn. *arXiv preprint arXiv:2306.08640*, 2023.
- [27] Luyu Gao, Aman Madaan, Shuyan Zhou, Uri Alon, Pengfei Liu, Yiming Yang, Jamie Callan, and Graham Neubig. Pal: Program-aided language models. *arXiv preprint arXiv:2211.10435*, 2022.
- [28] 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 arXiv:2304.15010*, 2023.
- [29] Tao Gong, Chengqi Lyu, Shilong Zhang, Yudong Wang, Miao Zheng, Qian Zhao, Kuikun Liu, Wenwei Zhang, Ping Luo, and Kai Chen. Multimodal-gpt: A vision and language model for dialogue with humans, 2023.
- [30] Google. Bard. <https://bard.google.com>, 2023. Accessed: 2023-07-17.
- [31] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913, 2017.
- [32] Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022.
- [33] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *CVPR*, 2019.
- [34] Huggingface. Transformers agent. https://huggingface.co/docs/transformers/transformers_agents, 2023. Accessed: 2023-07-20.
- [35] Geewook Kim, Teakgyu Hong, Moonbin Yim, JeongYeon Nam, Jinyoung Park, JinYeong Yim, Wonseok Hwang, Sangdoo Yun, Dongyoon Han, and Seunghyun Park. Ocr-free document understanding transformer. In *European Conference on Computer Vision*, pages 498–517. Springer, 2022.
- [36] Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *ICML*, 2021.
- [37] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *IJCV*, 2017.

- [38] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*, 2019.
- [39] 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 arXiv:2306.05425*, 2023.
- [40] 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 arXiv:2305.03726*, 2023.
- [41] Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. Seed-bench: Benchmarking multimodal llms with generative comprehension, 2023.
- [42] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023.
- [43] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR, 2022.
- [44] Xiujun Li, Xi Yin, Chunyuan Li, Xiaowei Hu, Pengchuan Zhang, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *ECCV*, 2020.
- [45] 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.
- [46] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*, 2023.
- [47] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*, 2023.
- [48] Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. Gpteval: Nlg evaluation using gpt-4 with better human alignment. *arXiv preprint arXiv:2303.16634*, 2023.
- [49] 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 arXiv:2307.06281*, 2023.
- [50] Yuliang Liu, Zhang Li, Hongliang Li, Wenwen Yu, Mingxin Huang, Dezheng Peng, Mingyu Liu, Mingrui Chen, Chunyuan Li, Lianwen Jin, et al. On the hidden mystery of ocr in large multimodal models. *arXiv preprint arXiv:2305.07895*, 2023.
- [51] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *NeurIPS*, 2019.
- [52] Timo Lüddecke and Alexander Ecker. Image segmentation using text and image prompts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7086–7096, 2022.
- [53] Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *CVPR*, pages 3195–3204, 2019.
- [54] OpenAI. Gpt-4 technical report, 2023.
- [55] Vicente Ordonez, Girish Kulkarni, and Tamara L Berg. Im2text: Describing images using 1 million captioned photographs. In *NeurIPS*, 2011.
- [56] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022.
- [57] Baolin Peng, Chunyuan Li, Pengcheng He, Michel Galley, and Jianfeng Gao. Instruction tuning with gpt-4. *arXiv preprint arXiv:2304.03277*, 2023.
- [58] Thomas Politzer. Vision is our dominant sense. <https://www.brainline.org/article/vision-our-dominant-sense>. Accessed: 2023-05-20.

- [59] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*, 2021.
- [60] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavy, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*, pages 28492–28518. PMLR, 2023.
- [61] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021.
- [62] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *ACL*, 2018.
- [63] Yongliang Shen, Kaitao Song, Xu Tan, Dongsheng Li, Weiming Lu, and Yuetong Zhuang. Hugginggpt: Solving ai tasks with chatgpt and its friends in huggingface. *arXiv preprint arXiv:2303.17580*, 2023.
- [64] Oleksii Sidorov, Ronghang Hu, Marcus Rohrbach, and Amanpreet Singh. Textcaps: a dataset for image captioning with reading comprehension. In *ECCV*, pages 742–758, 2020.
- [65] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8317–8326, 2019.
- [66] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.
- [67] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.
- [68] Maria Tsimpoukelli, Jacob Menick, Serkan Cabi, SM Eslami, Oriol Vinyals, and Felix Hill. Multimodal few-shot learning with frozen language models. *arXiv preprint arXiv:2106.13884*, 2021.
- [69] 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 arXiv:2205.14100*, 2022.
- [70] Wenhui 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 arXiv:2305.11175*, 2023.
- [71] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhdadi Bagheri, and Ronald M Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2097–2106, 2017.
- [72] Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. Simvlm: Simple visual language model pretraining with weak supervision. In *ICLR*, 2022.
- [73] 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 arXiv:2306.09265*, 2023.
- [74] Zhengyuan Yang, Zhe Gan, Jianfeng Wang, Xiaowei Hu, Faisal Ahmed, Zicheng Liu, Yumao Lu, and Lijuan Wang. Unitab: Unifying text and box outputs for grounded vision-language modeling. In *European Conference on Computer Vision*, pages 521–539. Springer, 2022.
- [75] Zhengyuan Yang, Zhe Gan, Jianfeng Wang, Xiaowei Hu, Yumao Lu, Zicheng Liu, and Lijuan Wang. An empirical study of gpt-3 for few-shot knowledge-based vqa. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 3081–3089, 2022.
- [76] Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Ehsan Azarnasab, Faisal Ahmed, Zicheng Liu, Ce Liu, Michael Zeng, and Lijuan Wang. Mm-react: Prompting chatgpt for multimodal reasoning and action. *arXiv preprint arXiv:2303.11381*, 2023.

- [77] Zhengyuan Yang, Yijuan Lu, Jianfeng Wang, Xi Yin, Dinei Florencio, Lijuan Wang, Cha Zhang, Lei Zhang, and Jiebo Luo. Tap: Text-aware pre-training for text-vqa and text-caption. In *CVPR*, pages 8751–8761, 2021.
- [78] Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, et al. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*, 2023.
- [79] Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. From recognition to cognition: Visual commonsense reasoning. In *CVPR*, pages 6720–6731, 2019.
- [80] Andy Zeng, Adrian Wong, Stefan Welker, Krzysztof Choromanski, Federico Tombari, Aveek Purohit, Michael Ryoo, Vikas Sindhwani, Johnny Lee, Vincent Vanhoucke, et al. Socratic models: Composing zero-shot multimodal reasoning with language. *arXiv preprint arXiv:2204.00598*, 2022.
- [81] 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 arXiv:2307.02469*, 2023.
- [82] Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuhui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022.
- [83] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric. P Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena, 2023.
- [84] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023.
- [85] Wanrong Zhu, Jack Hessel, Anas Awadalla, Samir Yitzhak Gadre, Jesse Dodge, Alex Fang, Youngjae Yu, Ludwig Schmidt, William Yang Wang, and Yejin Choi. Multimodal c4: An open, billion-scale corpus of images interleaved with text. *arXiv preprint arXiv:2304.06939*, 2023.