



MM-Vet v2: A Challenging Benchmark to Evaluate Large Multimodal Models for Integrated Capabilities

Weihao Yu^{1*} Zhengyuan Yang^{2*} Lingfeng Ren^{1*} Linjie Li² Jianfeng Wang²

Kevin Lin² Chung-Ching Lin² Zicheng Liu³ Lijuan Wang^{2†} Xinchao Wang^{1†}

¹National University of Singapore ²Microsoft ³Advanced Micro Devices

{weihaoyu,lingfengren}@u.nus.edu zicliu@outlook.com xinchao@nus.edu.sg

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

⌚ Code & data: <https://github.com/yuweihao/MM-Vet>

👉 Online evaluator: https://huggingface.co/spaces/whyu/MM-Vet-v2_Evaluator

📊 Leaderboard: <https://paperswithcode.com/sota/visual-question-answering-on-mm-vet-v2>

Abstract

MM-Vet, with open-ended vision-language questions targeting at evaluating integrated capabilities, has become one of the most popular benchmarks for large multimodal model evaluation. MM-Vet assesses six core vision-language (VL) capabilities: recognition, knowledge, spatial awareness, language generation, OCR, and math. However, its question format is restricted to single image-text pairs, lacking the interleaved image and text sequences prevalent in real-world scenarios. To address this limitation, we introduce MM-Vet v2, which includes a new VL capability called “image-text sequence understanding”, evaluating models’ ability to process VL sequences. Furthermore, we maintain the high quality of evaluation samples while further expanding the evaluation set size. Using MM-Vet v2 to benchmark large multimodal models, we found that Claude 3.5 Sonnet is the best model with a score of 71.8, slightly outperforming GPT-4o which scored 71.0. Among open-weight models, InternVL2-Llama3-76B leads with a score of 68.4.

1 Introduction

Large multimodal models (LMMs) [17, 20, 4] evolve rapidly, demonstrating emergent abilities to solve complex tasks [29] that required multiple integrated capabilities, such as GUI navigation [28, 30, 21], screenshot to code [23, 11], video understanding [15, 33], etc. To comprehensively evaluate LMMs, multiple benchmarks have been proposed, such as MME [10], MMBench [18], SEED-Bench [14], MMMU [32], and MM-Vet [31]. Notably, MM-Vet is designed to evaluate LMMs from a capability integration perspective, defining tasks based on their required core capabilities. The benchmark accepts open-ended responses and takes GPT-4 [19] to score model predictions, which better align with real-world scenarios. Such effective evaluation designs make MM-Vet widely utilized as a standard benchmark for LMM evaluation, as indicated by its leaderboard³.

Despite its popularity, MM-Vet and other concurrent evaluation benchmarks [10, 18, 14] are beginning to fall short in assessing the more advanced capabilities that have emerged in the latest LMMs, such as GPT-4V [20] and its successors [24, 8, 17, 7, 22, 5]. Specifically, one major limitation is the question format. The questions in MM-Vet are limited to a single image-text pair, lacking the capability to handle interleaved image and text sequences. This design choice was natural at the time of prior studies [31], given that most LMMs only supported single image inputs. However, the ability to process arbitrarily interleaved image-text sequences is crucial for advanced LMMs and should be included in LMM evaluation.

*Equal contribution.

†Corresponding authors.

³MM-Vet leaderboard: <https://paperswithcode.com/sota/visual-question-answering-on-mm-vet>

MM-Vet v2 examples

Prompt: How many feet do these animals have in total?

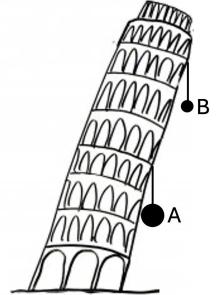


Ground truth: 10

Required capabilities: Recognition, Knowledge
Note: GPT-4V wrongly answers "14" including the rabbit in the mirror.

(a)

Prompt: As shown in the image, two iron balls are hanging on the Leaning Tower of Pisa, ball A weighs 20kg, and ball B weighs 5kg. If the ropes hanging them are cut at the same time and air resistance is ignored, which iron ball will land first?



Ground truth: A

Required capabilities: Recognition, OCR, Spatial Awareness, Knowledge
Note: GPT-4V wrongly answers "the same".

(b)

Prompt: Explain the story shown in the images below.



Ground truth: The images show a reel on Instagram where a user humorously reveals the extravagant sum they've spent on their dog by employing Biden's quotes of hefty figures for comedic effect. The first image shows a man holding his dog with the caption "When someone asks me how much I have spent on my dog". The next three images show a clip of Joe Biden saying "700 billion," "700 billion and a trillion 300 million" and "700 billion and a trillion 300 billion dollars", respectively.

Required capabilities: Recognition, OCR, Knowledge, **Sequence Understanding**, Language Generation

(c)

Prompt: Compare Image 1 and image 2, tell me about 3 differences between image 1 and image 2.
Image 1



Ground truth: 1) In image 1, the color of the coral reef on the right is blue while in Image 2, the color is red.
2) In image 1, there is a small red coral reef in the middle. In image 2, there is a green plant in the middle.
3) In image 1, the turtle's head is facing downwards. In image 2 the turtle's head is facing upwards.

Required capabilities: Recognition, **Sequence Understanding**, Language Generation

(d)

Figure 1: Four examples from MM-Vet v2. Compared with MM-Vet [31], MM-Vet v2 introduces more high-quality evaluation samples (e.g., (a) and (b)), and the ones with the new capability of image-text sequence understanding (e.g., (c) and (d)).

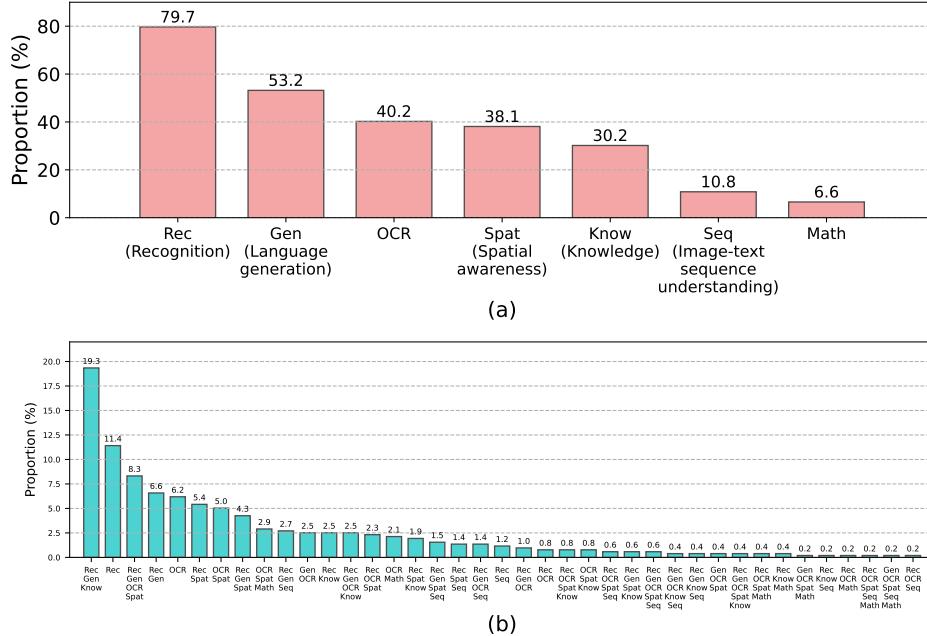


Figure 2: Proportions of (a) core capabilities and (b) capability integrations on MM-Vet v2.

In addition to the six core capabilities defined in MM-Vet, we introduce an additional capability: “image-text sequence understanding.” This measures the LMMs’ ability to process image-text sequential data, as shown in Figure 1. For example, to complete the task in Figure 1(c), the LMM needs to understand the text question followed by multiple sequential images. This task requires image-text sequence understanding alongside the capabilities of recognition, OCR, knowledge, and language generation defined in MM-Vet. The image-text sequence can contain multiple images of drastically different usages, such as multiple video frames in Figure 1(c) for temporal understanding, and spot the difference challenge in Figure 1(d) for image comparison. These capabilities to comprehend arbitrarily interleaved image-text sequences is one fundamental step towards stronger general intelligence.

Another limitation is the size of the evaluation set. MM-Vet only has 217 evaluation samples due to the difficulty of high-quality data collection. We aim to maintain the same high quality while further expanding the evaluation set size. We break down the challenge of creating high-quality evaluation samples into two steps: generating high-quality questions and producing reference answers. For the question generation, we find it difficult for crowd-sourcing workers to propose a meaningful and complicated question that covers various scenarios, *e.g.*, in Figure 1 (a,b). Instead, we have researchers design and collect 517 questions covering various scenarios from daily life to expert/industry applications, extended from an exploratory report on GPT-4V [29]. For straightforward questions that can be answered in a few words, our experts directly annotate the reference answer. For questions that need long paragraphs to answer, we first employ the GPT-4V [20] to draft the response. Next, our experts correct draft if there is any error and then rephrase it into the final ground truth.

After developing MM-Vet v2, we evaluate multiple advanced LMMs. Claude 3.5 Sonnet [1] achieves the highest performance with a score of 71.8, slightly surpassing GPT-4o [2] by 0.8 points. Notably, InternVL2-Llama3-76B [3], an open-weight model, also delivers a very competitive score of 68.4.

2 Dataset and evaluator

The same as MM-Vet [31], we aim to build a high-quality evaluation set for large multimodal models. MM-Vet [31] defines six core vision-language capabilities, including recognition (Rec), knowledge (Know), OCR, spatial awareness (Spat), language generation (Gen), and Math. MM-Vet’s question format is only an image-text pair, which obviously cannot measure the capability of processing sequential image and text data. To fill this gap, we introduce a new capability:

Table 1: MM-Vet v2’s few-shot prompt to evaluate model outputs using GPT-4 (gpt-4-0613), where \mathcal{Q} represents a sample’s question, \mathcal{G} denotes the ground truth and \mathcal{P} is the model output for the sample. Compared with MM-Vet, MM-Vet v2 adds $\langle\text{image}\rangle$ to represent image position in the question.

<p>Compare the ground truth and prediction from AI models, to give a correctness score for the prediction. $\langle\text{image}\rangle$ in the question indicates where an image is. $\langle\text{AND}\rangle$ in the ground truth means it is totally right only when all elements in the ground truth are present in the prediction, and $\langle\text{OR}\rangle$ 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.</p>				
Question Ground truth Prediction Correctness				
— — — —				
What is x in the equation? $\langle\text{image}\rangle$ -1 $\langle\text{AND}\rangle$ -5 $x = 3$ 0.0				
What is x in the equation? $\langle\text{image}\rangle$ -1 $\langle\text{AND}\rangle$ -5 $x = -1$ 0.5				
What is x in the equation? $\langle\text{image}\rangle$ -1 $\langle\text{AND}\rangle$ -5 $x = -5$ 0.5				
What is x in the equation? $\langle\text{image}\rangle$ -1 $\langle\text{AND}\rangle$ -5 $x = -5$ or 5 0.5				
What is x in the equation? $\langle\text{image}\rangle$ -1 $\langle\text{AND}\rangle$ -5 $x = -1$ or $x = -5$ 1.0				
Can you explain this meme? $\langle\text{image}\rangle$ 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? $\langle\text{image}\rangle$ 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}				

- **Image-text sequence understanding (Seq).** It refers to the capability to understand and reason the relationships among sequential image and text streaming data.

Then we carefully design 517 questions that require one or more capabilities, extended from MM-Vet [31] and an exploratory report based on [29]. Specifically, 218 questions are directly from MM-Vet [31], and the others are newly collected in various domains from daily life to expert/industry applications. For those questions that can be answered shortly, we annotated the ground truths directly. For questions requiring long text to respond, we first utilize GPT-4V [20] to generate the answer drafts, then we carefully correct the wrong information in the drafts and rephrase them into final ground truths. Because this new dataset is the extension of MM-Vet [31], we name it MM-Vet v2. See Figure 2 for the proportions of each core capability and capability integration.

Next, we design the evaluator for MM-Vet v2, based on that of MM-Vet [31]. Different from MM-Vet, the input format of MM-Vet v2 is not only an image-text pair, it can also be image-text sequences. To demonstrate the position of the image in the sequences, we utilize $\langle\text{image}\rangle$ to represent the image position. The few-shot prompt for MM-Vet v2 is shown in Table 1.

3 Experiments

3.1 Experiment settings

We evaluate two types of LMMs on our MM-Vet v2: (1) open-weight LMMs including OpenFlamingo [4, 6], Otter [13], LLaVA [17, 16], CogAgent [12], Emu2-Chat [25], InternLM-XComposer2 (IXC2)

Table 2: MM-Vet v2 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	MM-Vet	Rec	Gen	OCR	Spat	Know	Seq	Math	MM-Vet-v2
OpenFlamingo-9B [4, 6]	24.8	19.1	11.0	11.2	13.0	18.4	10.2	2.9	17.6±0.2
Otter-9B [13]	24.7	25.1	16.5	13.7	19.5	23.5	9.9	6.4	23.2±0.1
LLaVA-v1.5-7B [17]	34.2	30.4	21.5	21.7	25.4	23.6	9.5	6.0	28.3±0.2
LLaVA-v1.5-13B [17]	39.2	34.8	29.2	28.7	29.1	29.4	17.8	8.8	33.2±0.1
CogAgent-Chat [12]	40.5	33.6	29.2	35.3	27.3	33.9	24.7	5.9	34.7±0.2
Emu2-Chat [25]	45.5	37.9	28.8	35.1	35.1	37.9	18.7	10.8	38.0±0.1
IXC2-VL-7B [9]	52.3	40.7	37.0	42.5	37.0	39.3	5.2	36.1	42.5±0.3
CogVLM-Chat [27]	52.6	46.3	40.7	41.6	40.3	47.5	33.8	7.7	45.1±0.2
InternVL-Chat-V1-2 [8]	49.7	46.1	40.8	43.8	41.4	42.1	25.1	11.8	45.5±0.1
LLaVA-NeXT-34B [17]	60.2	49.3	48.9	52.8	48.3	49.9	18.5	35.1	50.9±0.1
InternVL-Chat-V1-5 [8]	62.8	52.0	48.9	51.0	49.3	48.2	37.6	16.5	51.5±0.2
InternVL2-40B [3]	61.8	63.6	63.9	64.3	60.0	60.9	48.5	45.3	63.8±0.2
InternVL2-Llama3-76B [3]	64.4	67.0	68.5	70.6	65.2	62.4	59.4	55.4	68.4±0.3
Claude 3 Opus [5]	58.6	53.5	57.6	60.7	50.0	51.3	46.1	42.9	55.8±0.2
Qwen-VL-Max [7]	66.6	51.7	51.1	60.2	49.0	52.5	27.3	60.2	55.8±0.2
Gemini Pro Vision [26]	63.1	54.3	50.8	61.5	55.8	50.4	45.4	43.5	57.2±0.2
GPT-4V [20]	67.7	63.1	67.1	73.6	65.8	53.4	62.2	68.3	66.3±0.2
Gemini 1.5 Pro [22]	65.8	64.4	64.7	75.1	65.9	56.6	63.9	61.5	66.9±0.2
GPT-4o [2]	69.3	67.5	70.5	78.0	68.8	63.8	74.3	77.6	71.0±0.2
Claude 3.5 Sonnet [1]	74.2	69.2	70.8	78.9	68.8	65.6	67.7	69.1	71.8±0.2

Table 3: MM-Vet v2 evaluation results on various LMMs regarding *capability integrations*. Due to space limitations, only the 16 integrations with the highest proportions are displayed. 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			Rec			Rec			Rec			Total				
	Rec	Gen	OCR	Rec													
	Gen	OCR	Rec	Rec	Gen	OCR	Rec	Gen	OCR	Rec	Gen	OCR	Rec				
Know	Rec	Spat	Gen	OCR	Rec	Spat	Know	Rec	Spat	Gen	Know	Rec	Spat	Math	Know		
OpenFlamingo-9B [4, 6]	14.9	42.7	6.1	18.3	20.2	19.3	14.6	8.1	6.7	13.3	0.0	46.2	2.8	16.7	0.0	30.4	17.6±0.2
Otter-9B [13]	20.5	48.2	5.0	38.8	27.2	47.8	17.3	10.9	14.5	8.4	0.0	23.8	17.7	12.5	0.0	58.0	23.2±0.1
LLaVA-v1.5-7B [17]	23.7	64.2	24.5	36.5	38.4	58.2	19.6	10.9	13.3	2.9	0.5	15.4	24.6	12.5	0.0	37.0	28.3±0.2
LLaVA-v1.5-13B [17]	26.5	60.7	33.9	55.5	49.9	43.2	28.5	12.8	13.3	16.9	0.0	26.9	43.4	20.8	0.0	44.6	33.2±0.1
CogAgent-Chat [12]	30.1	59.7	29.3	33.6	64.9	38.9	44.2	3.5	13.3	26.6	26.5	38.5	60.2	8.3	0.0	26.4	34.7±0.2
Emu2-Chat [25]	34.6	67.1	38.8	28.6	64.4	61.4	48.0	5.2	13.3	21.7	5.7	49.2	38.3	8.3	0.0	59.4	38.0±0.1
IXC2-VL-7B [9]	40.9	66.9	40.4	60.9	71.1	64.6	48.1	20.4	46.0	3.6	13.8	15.4	56.3	0.0	36.4	30.8	42.5±0.3
CogVLM-Chat [27]	44.4	72.2	36.9	46.1	53.6	59.6	55.7	18.7	10.8	32.7	22.2	37.7	61.2	33.3	9.1	65.0	45.1±0.2
InternVL-Chat-V1-2 [8]	39.6	69.9	46.9	57.9	75.7	71.8	48.1	22.2	13.3	32.3	18.6	29.2	64.2	16.7	9.1	35.2	45.5±0.1
LLaVA-NeXT-34B [17]	53.2	66.7	57.0	55.1	71.6	65.3	65.0	44.4	46.7	10.0	32.3	30.3	73.5	16.7	34.5	41.6	50.9±0.1
InternVL-Chat-V1-5 [8]	47.5	72.1	59.2	56.6	69.1	65.3	69.2	37.7	23.6	42.7	22.9	43.1	70.8	25.0	0.0	43.4	51.5±0.2
InternVL2-40B	57.8	77.7	70.0	75.0	81.6	75.0	62.3	59.9	53.3	57.9	54.6	38.9	86.5	25.0	40.0	66.8	63.8±0.2
InternVL2-Llama3-76B [3]	62.4	80.4	74.5	77.7	75.2	78.1	79.5	60.5	66.7	67.7	74.8	50.0	89.1	25.0	63.6	47.4	68.4±0.3
Claude 3 Opus [5]	53.9	63.8	67.5	72.1	76.4	45.0	64.8	34.1	46.7	46.9	57.4	29.2	62.5	25.0	50.0	29.0	55.8±0.2
Qwen-VL-Max [7]	52.6	71.2	53.5	59.2	76.8	64.6	76.7	41.9	66.0	33.4	68.6	55.4	79.1	8.3	63.6	15.0	55.8±0.2
Gemini Pro Vision [26]	45.4	67.3	51.3	59.1	83.8	76.1	76.7	48.6	52.0	39.9	50.6	74.6	67.4	33.3	51.8	40.0	57.2±0.2
GPT-4V [20]	56.4	71.0	82.4	77.6	85.5	63.1	80.4	62.6	86.0	62.7	64.2	37.7	80.0	41.7	63.6	35.0	66.3±0.2
Gemini 1.5 Pro [22]	54.3	71.0	79.0	70.3	86.6	77.1	73.1	44.4	73.3	77.1	78.9	74.6	76.0	41.7	63.6	45.0	66.9±0.2
GPT-4o [2]	62.4	71.1	79.1	79.5	89.9	65.5	82.5	54.1	99.3	75.7	73.5	46.8	86.2	25.0	63.6	72.6	71.0±0.2
Claude 3.5 Sonnet [1]	65.2	79.7	79.2	68.5	83.8	73.6	86.5	48.4	86.7	81.7	80.6	50.0	83.8	66.7	54.5	41.0	71.8±0.2

[9], CogVLM [27] and InternVL [8]; (2) Closed source models including Claude [5, 1], Qwen-VL Max [7], Gemini [26, 22], GPT-4V [20] and GPT-4o [2].

As illustrated in 1, for each sample, we complete the prompt template with its question, ground truth, and the output from a specific LMM. When this filled prompt is input into GPT-4, it generates a score ranging from 0 to 1 for each sample. Although the temperature is set to 0, we observe some variance in GPT-4’s outputs. To address this, we evaluate the outputs of the LLMs using GPT-4 five times. Due to space constraints, we present the average scores for capabilities and for some capability integrations, and include both the average and variance for the total score.

3.2 Results

The main results of the different methods are presented in Table 2 for each capability, and in Table 3 for 16 integrations with the highest proportions. Claude 3.5 Sonnet [1] and GPT-4o [2] are the leading models, achieving scores of 71.8 and 71.0, respectively. Claude 3.5 Sonnet [1] excels in recognition, language generation, OCR, spatial awareness, and knowledge. On the other hand, GPT-4o [2] surpasses in image-text sequence understanding and math. Among open-weight models, InternVL2-Llama3-76B stands out with a competitive performance score of 68.4.

4 Conclusion

In this paper, we aim to evaluate the integrated capabilities of large multimodal models and extend MM-Vet into MM-Vet v2 by introducing a new core capability: Image-text sequence understanding, which assesses the ability to process vision-language sequences. Additionally, we ensure the high quality of evaluation samples while expanding the evaluation set size. Using MM-Vet v2 to benchmark large multimodal models, we found that Claude 3.5 Sonnet achieved the highest score of 71.8, narrowly surpassing GPT-4o, which scored 71.0. Among open-weight models, InternVL2-Llama3-76B emerged as the leader with a score of 68.4.

References

- [1] Claude 3.5 sonnet. <https://www.anthropic.com/news/clause-3-5-sonnet>, 2024.
- [2] Hello gpt-4o. <https://openai.com/index/hello-gpt-4o/>, 2024.
- [3] Internvl2: Better than the best—expanding performance boundaries of open-source multimodal models with the progressive scaling strategy. <https://internvl.github.io/blog/2024-07-02-InternVL-2.0/>, 2024.
- [4] 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.
- [5] Anthropic. Claude 3 model card. 2024.
- [6] Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Shiori Sagawa, Jenia Jitsev, Simon Kornblith, Pang Wei Koh, Gabriel Ilharco, Mitchell Wortsman, and Ludwig Schmidt. Openflamingo: An open-source framework for training large autoregressive vision-language models. *arXiv preprint arXiv:2308.01390*, 2023.
- [7] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. 2023.
- [8] Zhe Chen, Jiannan Wu, Wenhui Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24185–24198, 2024.
- [9] Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Bin Wang, Linke Ouyang, Xilin Wei, Songyang Zhang, Haodong Duan, Maosong Cao, et al. Internlm-xcomposer2: Mastering free-form text-image composition and comprehension in vision-language large model. *arXiv preprint arXiv:2401.16420*, 2024.
- [10] 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.
- [11] Yucheng Han, Chi Zhang, Xin Chen, Xu Yang, Zhibin Wang, Gang Yu, Bin Fu, and Hanwang Zhang. Chartllama: A multimodal llm for chart understanding and generation. *arXiv preprint arXiv:2311.16483*, 2023.
- [12] Wenyi Hong, Weihua Wang, Qingsong Lv, Jiazhen Xu, Wenmeng Yu, Junhui Ji, Yan Wang, Zihan Wang, Yuxiao Dong, Ming Ding, et al. Cogagent: A visual language model for gui agents. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14281–14290, 2024.

- [13] 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.
- [14] Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. Seed-bench: Benchmarking multimodal llms with generative comprehension, 2023.
- [15] Kevin Lin, Faisal Ahmed, Linjie Li, Chung-Ching Lin, Ehsan Azarnasab, Zhengyuan Yang, Jianfeng Wang, Lin Liang, Zicheng Liu, Yumao Lu, et al. Mm-vid: Advancing video understanding with gpt-4v (ision). *arXiv preprint arXiv:2310.19773*, 2023.
- [16] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306, 2024.
- [17] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*, 2023.
- [18] 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.
- [19] OpenAI. Gpt-4 technical report, 2023.
- [20] OpenAI. Gpt-4v(ision) system card. 2023.
- [21] Kevin Qinghong Lin, Linjie Li, Difei Gao, Qinchen WU, Mingyi Yan, Zhengyuan Yang, Lijuan Wang, and Mike Zheng Shou. Videogui: A benchmark for gui automation from instructional videos. *arXiv e-prints*, pages arXiv–2406, 2024.
- [22] Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy Lillicrap, Jean-baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024.
- [23] Chenglei Si, Yanzhe Zhang, Zhengyuan Yang, Ruibo Liu, and Diyi Yang. Design2code: How far are we from automating front-end engineering? *arXiv preprint arXiv:2403.03163*, 2024.
- [24] Quan Sun, Yufeng Cui, Xiaosong Zhang, Fan Zhang, Qiying Yu, Zhengxiong Luo, Yueze Wang, Yongming Rao, Jingjing Liu, Tiejun Huang, et al. Generative multimodal models are in-context learners. *arXiv preprint arXiv:2312.13286*, 2023.
- [25] Quan Sun, Yufeng Cui, Xiaosong Zhang, Fan Zhang, Qiying Yu, Yueze Wang, Yongming Rao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. Generative multimodal models are in-context learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14398–14409, 2024.
- [26] Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.
- [27] Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, et al. Cogvilm: Visual expert for pretrained language models. *arXiv preprint arXiv:2311.03079*, 2023.
- [28] An Yan, Zhengyuan Yang, Wanrong Zhu, Kevin Lin, Linjie Li, Jianfeng Wang, Jianwei Yang, Yiwu Zhong, Julian McAuley, Jianfeng Gao, et al. Gpt-4v in wonderland: Large multimodal models for zero-shot smartphone gui navigation. *arXiv preprint arXiv:2311.07562*, 2023.
- [29] Zhengyuan Yang, Linjie Li, Kevin Lin, Jianfeng Wang, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. The dawn of lmms: Preliminary explorations with gpt-4v (ision). *arXiv preprint arXiv:2309.17421*, 2023.
- [30] Keen You, Haotian Zhang, Eldon Schoop, Floris Weers, Amanda Swearngin, Jeffrey Nichols, Yinfei Yang, and Zhe Gan. Ferret-ui: Grounded mobile ui understanding with multimodal llms. *arXiv preprint arXiv:2404.05719*, 2024.
- [31] Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*, 2023.

- [32] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9556–9567, 2024.
- [33] Chaoyi Zhang, Kevin Lin, Zhengyuan Yang, Jianfeng Wang, Linjie Li, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. Mm-narrator: Narrating long-form videos with multimodal in-context learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13647–13657, 2024.