Improved Baselines with Visual Instruction Tuning

Haotian Liu[1] **Chunyuan Li**[2] **Yuheng Li**[1] **Yong Jae Lee**[1]
[1]University of Wisconsin–Madison [2]Microsoft Research, Redmond
https://llava-vl.github.io

Abstract

Large multimodal models (LMM) have recently shown encouraging progress with visual instruction tuning. In this note, we show that the fully-connected vision-language cross-modal connector in LLaVA is surprisingly powerful and data-efficient. With simple modifications to LLaVA, namely, using CLIP-ViT-L-336px with an MLP projection and adding academic-task-oriented VQA data with simple response formatting prompts, we establish stronger baselines that achieve state-of-the-art across 11 benchmarks. Our final 13B checkpoint uses merely 1.2M publicly available data, and finishes full training in ∼1 day on a single 8-A100 node. We hope this can make state-of-the-art LMM research more accessible. Code and model will be publicly available.

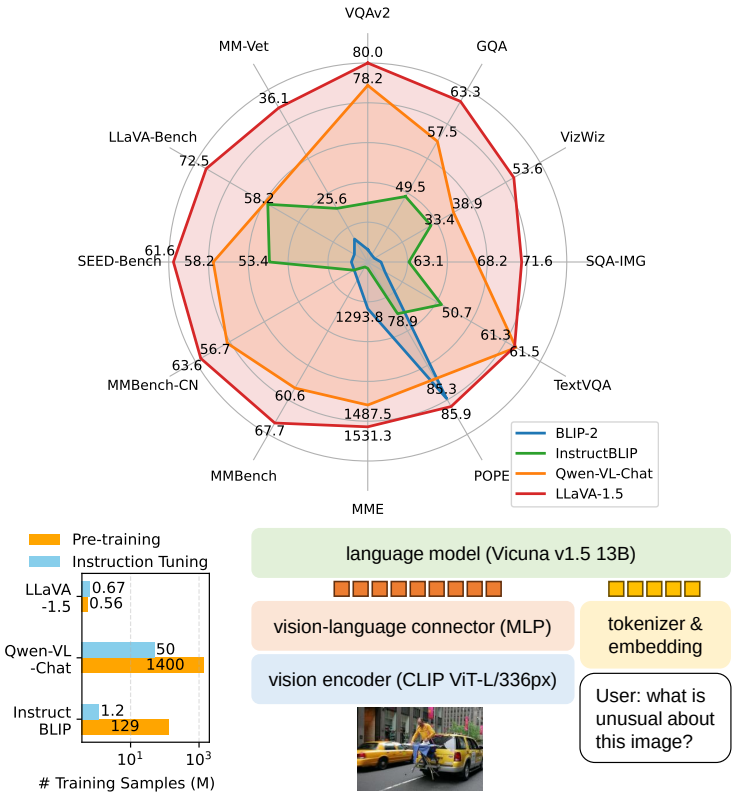

Figure 1: **LLaVA-1.5** achieves SoTA on a broad range of 11 tasks (Top), with high training sample efficiency (Left) and simple modifications to LLaVA (Right): an MLP connector and including academic-task-oriented data with response formatting prompts.

Workshop on Instruction Tuning and Instruction Following at NeurIPS 2023.

1 Introduction

Large multimodal models (LMMs) have become increasingly popular in the research community, as they are the key building blocks towards general-purpose assistants [21, 34, 1]. Recent studies on LMMs are converging on a central concept known as visual instruction tuning [27]. The results are promising, *e.g.* LLaVA [27] and MiniGPT-4 [46] demonstrate impressive results on natural instruction-following and visual reasoning capabilities. To better understand the capability of LMMs, multiple benchmarks [28, 40, 20, 25, 11] have been proposed. Recent works further demonstrate improved performance by scaling up the pretraining data [9, 2], instruction-following data [42, 43, 19, 9], visual encoders [2], or langauge models [30], respectively. The LLaVA architecture is also leveraged in different downstream tasks and domains, including region-level [6, 41] and pixel-level [18] understanding, biomedical assistants [22], image generation [3], adversarial studies [4, 44].

This note establishes stronger and more feasible baselines built upon the LLaVA framework. We report that two simple improvements, namely, an MLP cross-modal connector and incorporating academic task related data such as VQA, are orthogonal to the framework of LLaVA, and when used with LLaVA, lead to better multimodal understanding capabilities. In contrast to InstructBLIP [9] or Qwen-VL [2], which trains specially designed visual resamplers on hundreds of millions or even billions of image-text paired data, LLaVA uses the simplest architecture design for LMMs and requires only training a simple fully-connected projection layer on merely 600K image-text pairs. Our final model can finish training in $\sim$1 day on a single 8-A100 machine and achieves state-of-the-art results on a wide range of benchmarks. Moreover, unlike Qwen-VL [2] that includes in-house data in training, LLaVA utilizes only publicly available data. We hope these improved and easily-reproducible baselines will provide a reference for future research in open-source LMM.

2 Background

Instruction-following LMM. Common architectures include a pre-trained visual backbone to encode visual features, a pre-trained large language model (LLM) to comprehend the user instructions and produce responses, and a vision-language cross-modal connector to align the vision encoder outputs to the language models. As shown in Fig. 1, LLaVA [27] is perhaps the simplest architecture for LMMs. Optionally, visual resamplers (*e.g.* Qformer [23]) are used to reduce the number of visual patches [46, 9, 2]. Training an instruction-following LMM usually follows a two-stage protocol. First, the vision-language alignment pretraining stage leverages image-text pairs to align the visual features with the language model's word embedding space. Earlier works utilize relatively few image-text pairs (*e.g.* $\sim$600K [27] or $\sim$6M [46]), while some recent works pretrain the vision-language connector for a specific language model on a large amount of image-text pairs (*e.g.* 129M [9] and 1.4B [2]), to maximize the LMM's performance. Second, the visual instruction tuning stage tunes the model on visual instructions, to enable the model to follow users' diverse requests on instructions that involve the visual contents.

Multimodal instruction-following data. In NLP, studies show that the quality of instruction-following data largely affects the capability of the resulting instruction-following models [45]. For visual instruction tuning, LLaVA [27] is the pioneer to leverage text-only GPT-4 to expand the existing COCO [26] bounding box and caption dataset to a multimodal instruction-following dataset that contains three types of instruction-following data: conversational-style QA, detailed description, and complex reasoning. LLaVA's pipeline has been employed to expand to textual understanding [42], million-scales [43], and region-level conversations [6]. InstructBLIP [9] incorporates academic-task-oriented VQA datasets to further enhance the model's visual capabilities. Conversely, [5] identifies that such naive data merging can result in the models that tend to overfit to VQA datasets and thus are inability to participate in natural conversations. The authors further proposes to leverage the LLaVA pipeline to convert VQA datasets to a conversational style. While this proves effective for training, it introduces added complexities in data scaling.

3 Improved Baselines of LLaVA

Overview. As the initial work of visual instruction tuning, LLaVA has showcased commendable proficiency in visual reasoning capabilities, surpassing even more recent models on diverse benchmarks

	Method	LLM	Res.	GQA	MME	MM-Vet
	InstructBLIP	14B	224	49.5	1212.8	25.6
	Only using a subset of InstructBLIP training data					
0	**LLaVA**	7B	224	–	502.8	23.8
1	+VQA-v2	7B	224	47.0	1197.0	27.7
2	+Format prompt	7B	224	46.8	1323.8	26.3
3	+MLP VL connector	7B	224	47.3	1355.2	27.8
4	+OKVQA/OCR	7B	224	50.0	1377.6	29.6
	Additional scaling					
5	+Region-level VQA	7B	224	50.3	1426.5	30.8
6	+Scale up resolution	7B	336	51.4	1450	30.3
7	+GQA	7B	336	62.0*	1469.2	30.7
8	+ShareGPT	7B	336	62.0*	1510.7	31.1
9	+Scale up LLM	13B	336	**63.3***	**1531.3**	**36.1**

Table 1: **Scaling results** on data, model, and resolution. We choose to conduct experiments on GQA [14], MME [11], and MM-Vet [40] to examine the representative capabilities of VQA with short answers, VQA with output formatting, and natural visual conversations, respectively. *Training images of GQA were observed during training.

Input Image		Normal prompt	Response formatting prompt	Ambiguous prompt
	query	What is the color of the shirt that the man is wearing?	What is the color of the shirt that the man is wearing? **Answer the question using a single word or phrase.**	Q: What is the color of the shirt that the man is wearing? A:
	answer	The man is wearing a yellow shirt.	Yellow.	The man is wearing a yellow shirt.

Figure 2: Comparison of how different prompt regularizes the output format. The results are obtained zero-shot directly after LLaVA undergoes the first-stage vision-language alignment pretraining, without any visual instruction tuning.

for real-life visual instruction-following tasks, while only falling short on academic benchmarks that typically require short-form answers (*e.g.* single-word). The latter was attributed to the fact that LLaVA has not been pretrained on large-scale data, as other approaches do. In this note, we first study the scaling effect of data, models and input image resolution on a selection of three datasets in Table 1, and then compare the final model against existing LMMs on a diverse set of 12 benchmarks in Table 2. We show that the LLaVA's architecture is powerful and data-efficient for visual instruction tuning, and achieves the best performance using significantly less compute and training data than all other methods.

Response formatting prompts. We find that the inability [5] to balance between short- and long-form VQA for approaches like InstructBLIP [9] is mainly due to the following reasons. First, *ambiguous prompts on the response format*. For example, *Q: {Question} A: {Answer}*. Such prompts do not clearly indicate the desirable output format, and can overfit an LLM behavorially to short-form answers even for natural visual conversations. Second, *not finetuning the LLM*. The first issue is worsened by InstructBLIP only finetuning the Qformer for instruction-tuning. It requires the Qformer's visual output tokens to control the length of the LLM's output to be either long-form or short-form, as in prefix tuning [24], but Qformer may lack the capability of properly doing so, due to its limited capacity compared with LLMs like LLaMA. See Fig. 2 for an qualitative example.

To address this, we propose to use a single response formatting prompt that clearly indicates the output format, to be appended at the end of VQA questions when promoting short answers: *Answer the question using a single word or phrase*. We empirically show that when LLM is *finetuned* with such prompts, LLaVA is able to properly adjust the output format according to the user's instructions, and does not require additional processing of the VQA data using ChatGPT [5], which further enables scaling to various data sources. As shown in Table 1, by merely including VQAv2 [12] in training, LLaVA's performance on MME significantly improves (1323.8 *vs* 502.8) and outperforms InstructBLIP by 111 points.

MLP vision-language connector. Inspired by the improved performance in self-supervised learning by changing from a linear projection to an MLP [7, 8], we find that improving the vision-language

Method	LLM	Res.	PT	IT	VQA^{v2}	GQA	VW	SQA^{I}	VQA^{T}	POPE	MME	MMB	MMB^{CN}	SEED	LLaVA^{W}	MMVet
BLIP-2	13B	224	129M	-	41.0	41	19.6	61	42.5	85.3	1293.8	–	–	46.4	38.1	22.4
InstructBLIP	7B	224	129M	1.2M	–	49.2	34.5	60.5	50.1	–	–	36	23.7	53.4	60.9	26.2
InstructBLIP	13B	224	129M	1.2M	–	49.5	33.4	63.1	50.7	78.9	1212.8	–	–	–	58.2	25.6
IDEFICS-9B	7B	224	353M	1M	50.9	38.4	35.5	–	25.9	–	–	48.2	25.2	–	–	–
IDEFICS-80B	65B	224	353M	1M	60.0	45.2	36.0	–	30.9	–	–	54.5	38.1	–	–	–
Shikra	13B	224	600K	5.5M	77.4*	–	–	–	–	–	–	58.8	–	–	–	–
Qwen-VL	7B	448	1.4B†	50M†	78.8*	59.3*	35.2	67.1	**63.8**	–	–	38.2	7.4	–	–	–
Qwen-VL-Chat	7B	448	1.4B†	50M†	78.2*	57.5*	38.9	68.2	61.5	–	1487.5	60.6	56.7	58.2	–	–
LLaVA-1.5	7B	336	**558K**	**665K**	78.5*	62.0*	50.0	66.8	58.2	**85.9**	1510.7	64.3	58.3	58.6	65.4	31.1
LLaVA-1.5	13B	336	**558K**	**665K**	**80.0***	**63.3***	**53.6**	**71.6**	61.3	**85.9**	**1531.3**	**67.7**	**63.6**	**61.6**	**72.5**	**36.1**

Table 2: **Comparison with SoTA methods on 12 benchmarks.** LLaVA achieves the best performance on 11/12 benchmarks, and ranks the second on the other. Res, PT, IT indicate input image resolution, the number of samples in pretraining and instruction tuning stage, respectively. Benchmark names are abbreviated due to space limits. VQA-v2 [12]; GQA [14]; VW: VisWiz [13]; SQA^{I}: ScienceQA-IMG [29]; VQA^{T}: TextVQA [39]; POPE [25]; MME [11]; MMB: MMBench [28]; MMB^{CN}: MMBench-Chinese [28]; SEED: SEED-Bench [20]; LLaVA^{W}: LLaVA-Bench (In-the-Wild) [27]; MM-Vet [40]. *The training images of the datasets are observed during training. †Includes in-house data that is not publicly available. Qwen-VL uses Qwen, and other methods use Vicuna.

connector's representation power with a two-layer MLP can improve LLaVA's multimodal capabilities, compared with the original linear projection design.

Academic task oriented data. We further include additional academic-task-oriented VQA datasets for VQA, OCR, and region-level perception, to enhance the model's capabilities in various ways, as shown in Table 1. We first include four additional datasets that are used in InstructBLIP: open-knowledge VQA (OKVQA [32], A-OKVQA [36]) and OCR (OCRVQA [33], TextCaps [38]). A-OKVQA is converted to multiple choice questions and a specific response formatting prompt is used: *Answer with the option's letter from the given choices directly*. With only a subset of the datasets InstructBLIP uses, LLaVA already surpasses it on all three tasks in Table 1, suggesting LLaVA's effective design. Furthermore, we find further adding region-level VQA datasets (Visual Genome [17], RefCOCO [16, 31]) improves the model's capability of localizing fine-grained visual details.

Additional scaling. We further scale up the input image resolution to allows LLM to clearly "see" the details of images, and add the GQA dataset as an additional visual knowledge source. We also incorporate ShareGPT [37] data and scale up the LLM to 13B as in [2, 30, 6]. Results on MM-Vet shows the most significant improvement when scaling the LLM to 13B, suggesting the importance of the base LLM's capability for visual conversations. We denote the final model with all the modifications as LLaVA-1.5 (the last two rows in Table 1), which achieves an impressive performance that significantly outperforms the original LLaVA [27].

4 Discussion

Comparison with SoTA. We benchmark LLaVA-1.5 on a wide range of academic VQA benchmarks and recent benchmarks specifically proposed for instruction-following LMMs, totalling 12 benchmarks. We show that it achieves the best performance across 11 out of 12 benchmarks, despite using magnitudes smaller pretraining and instruction tuning data compared with other methods [9, 2]. It is encouraging that *LLaVA-1.5 achieves the best performance with the simplest architecture, academic compute and public datasets, and yields a fully-reproducible and affordable baseline for future research*. The results also suggest that visual instruction tuning plays a more important role in improving an LMM's capabilities than pretraining, and raises questions upon the common belief that LMMs require significant amount of vision-language alignment pretraining [9, 23, 2], despite that the vision encoders (*e.g.* CLIP [35], OpenCLIP [15], EVA-CLIP [10], *etc.*) are already pretrained on web-scale image-text paired dataset. This also makes us rethink the benefits of the vision samplers and the necessity of the additional large-scale pretraining, in terms of multimodal instruction-following capabilities.

Visual input example, Tricky Question:	
User	If there are factual errors in the questions, point it out; if not, proceed answering the question. What's happening in the desert?
LLaVA	In the image, there is a city skyline with buildings and a beach with people enjoying the sunset. The scene is set in a desert, which is unusual for a beach setting.
LLaVA-1.5	There are no deserts in the image. The image features a beach with palm trees, a city skyline, and a large body of water.

Table 3: LLaVA-1.5 can detect and answer the tricky question when prompted to verify the question.

Visual input example, Constrained JSON Output:	
User	Please read the text in this image and return the information in the following JSON format (note xxx is placeholder, if the information is not available in the image, put "N/A" instead). {"class": xxx, "DLN": xxx, "DOB": xxx, "Name": xxx, "Address": xxx, "EXP": xxx, "ISS": xxx, "SEX": xxx, "HGT": xxx, "WGT": xxx, "EYES": xxx, "HAIR": xxx, "DONOR": xxx}
GPT-4V	{"class": "C", "DLN": "1234568", "DOB": "08/31/1977", "Name": "Ima Cardholder", "Address": "2570 24th Street, Anytown, CA 95818", "EXP": "08/31/2014", "ISS": "08/31/2011", "SEX": "F", "HGT": "5-05", "WGT": "125 lbs", "EYES": "BRN", "HAIR": "BRN", "DONOR": "VETERAN"}
LLaVA	{"class": "Driver License", "DLN": "CA123 456 789", "DOB": "01/01/1990", "Name": "Amy", "Address": "Anytown, CA", "EXP": "08/31/2023", "ISS": "Veteran", "SEX": "F", "HGT": "5'5", "WGT": "130", "EYES": "Brown", "HAIR": "Brown", "DONOR": "Unknown"}
LLaVA-1.5	{"class": "C", "DLN": "1132456789", "DOB": "08/23/1971", "Name": "Ima Cardholder", "Address": "Anytown, State, Zip", "EXP": "08/23/2014", "ISS": "California", "SEX": "F", "HGT": "5'5", "WGT": "123", "EYES": "Brown", "HAIR": "Brown", "DONOR": "N/A"}

Table 4: LLaVA-1.5 can extract information from the image and respond to the user's request following the required format, despite a few errors compared with GPT-4V.

Computational cost. For LLaVA-1.5, we use the same pretraining dataset of LCS-558K[1], and keep the training iterations and batch size roughly the same for instruction tuning as LLaVA [27]. Due to the increased image input resolution to 336px, the training of LLaVA-1.5 is $\sim 2\times$ as long as LLaVA: $\sim$6 hours of pretraining and $\sim$20 hours of visual instruction tuning, using $8\times$ A100s.

Zero-shot format instruction generalization. Although LLaVA-1.5 is only instruction-tuned with two format instructions: *short answer* and *multiple-choice*, it generalizes to other format instructions. First, VizWiz [13] requires the model to output "Unanswerable" when the provided content is

[1]LCS-558K: a subset of $\sim$558K image-text pairs from LAION-CC-SBU with BLIP captions, as used in LLaVA-Lightning series.

insufficient to answer the question, and our response format prompt (Table 6) effectively instructs the model to do so (11.1% → 67.8% on unanswerable questions). We additionally present qualitative examples on instructing LLaVA-1.5 to verify the tricky questions (Fig. 3) and respond in a constrained JSON format (Fig. 4).

Zero-shot multilingual capability. Though LLaVA-1.5 is *not* finetuned for multilingual multimodal instruction following *at all*, we find that it is capable of following multilingual instructions, partly due to the multilingual language instructions in ShareGPT [37]. We quantitatively evaluate the model's generalization capability to Chinese on MMBench-CN [28], where the questions of MMBench are converted to Chinese. Notably, LLaVA-1.5 outperforms Qwen-VL-Chat by 7.3% (63.6% vs 56.7%), despite Qwen-VL-Chat being finetuned on Chinese multimodal instructions while LLaVA-1.5 is not.

Limitations. Despite the promising results demonstrated by LLaVA-1.5, several limitations must be acknowledged. First, LLaVA utilizes full image patches, potentially prolonging each training iteration. While visual resamplers [23, 9, 2] reduce the number of visual patches in LLMs, they currently cannot achieve convergence as efficiently as LLaVA with a comparable amount of training data, due to more trainable parameters in the resamplers. The development of a sample-efficient visual resampler could pave the way for future scaling-up of instruction-following multimodal models. Second, LLaVA-1.5 is not yet capable of processing multiple images due to the lack of such instruction-following data, and the limit of the context length. Third, although LLaVA-1.5 exhibits proficiency in following complex instructions, its problem-solving capabilities can still be limited in certain domains, which could be improved with a more capable language model and with high-quality, targeted visual instruction tuning data. Finally, despite its significantly reduced propensity for hallucination, LLaVA is not exempt from producing hallucinations and occasionally disseminating misinformation, and should be used with caution in critical applications (*e.g.* medical).

Data	Size	Response formatting prompts
LLaVA [27]	158K	–
ShareGPT [37]	40K	–
VQAv2 [12]	83K	Answer the question using a single word or phrase.
GQA [14]	72K	
OKVQA [32]	9K	
OCRVQA [33]	80K	
A-OKVQA [36]	50K	Answer with the option's letter from the given choices directly.
TextCaps [38]	22K	Provide a one-sentence caption for the provided image.
RefCOCO [16, 31]	30K	*Randomly choose between the two formats* Provide a short description for this region.
VG [17]	86K	Provide the bounding box coordinate of the region this sentence describes.
Total	665K	

Table 5: **Instruction-following Data Mixture** of LLaVA-1.5.

Data	Response formatting prompts
LLaVA-Bench, MM-Vet	–
VQAv2, GQA, TextVQA, MME, POPE	Answer the question using a single word or phrase.
ScienceQA, MMBench, SEED-Bench	Answer with the option's letter from the given choices directly.
VizWiz	When the provided information is insufficient, respond with 'Unanswerable'. Answer the question using a single word or phrase.

Table 6: **Format Prompt** for evaluation.

Hyperparameter	Pretrain	Finetune
batch size	256	128
lr	1e-3	2e-5
lr schedule	cosine decay	
lr warmup ratio	0.03	
weight decay	0	
epoch	1	
optimizer	AdamW	
DeepSpeed stage	2	3

Table 7: **Hyperparameters** of LLaVA-1.5 are the same as the original LLaVA, except we halve the learning rate in pretraining as we use MLP instead of linear projection layer.

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *arXiv preprint arXiv:2204.14198*, 2022. 2

[2] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 2023. 2, 4, 6

[3] Kevin Black, Michael Janner, Yilun Du, Ilya Kostrikov, and Sergey Levine. Training diffusion models with reinforcement learning. *arXiv preprint arXiv:2305.13301*, 2023. 2

[4] Nicholas Carlini, Milad Nasr, Christopher A Choquette-Choo, Matthew Jagielski, Irena Gao, Anas Awadalla, Pang Wei Koh, Daphne Ippolito, Katherine Lee, Florian Tramer, et al. Are aligned neural networks adversarially aligned? *arXiv preprint arXiv:2306.15447*, 2023. 2

[5] Delong Chen, Jianfeng Liu, Wenliang Dai, and Baoyuan Wang. Visual instruction tuning with polite flamingo. *arXiv preprint arXiv:2307.01003*, 2023. 2, 3

[6] Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. Shikra: Unleashing multimodal llm's referential dialogue magic. *arXiv preprint arXiv:2306.15195*, 2023. 2, 4

[7] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, 2020. 3

[8] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020. 3

[9] 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. 2, 3, 4, 6

[10] 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. 4

[11] 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. 2, 3, 4

[12] 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. 3, 4, 6

[13] Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3608–3617, 2018. 4, 5

[14] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *CVPR*, 2019. 3, 4, 6

[15] Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. Openclip. July 2021. If you use this software, please cite it as below. 4

[16] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 787–798, 2014. 4, 6

[17] 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. *International journal of computer vision*, 123:32–73, 2017. 4, 6

[18] Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. Lisa: Reasoning segmentation via large language model. *arXiv preprint arXiv:2308.00692*, 2023. 2

[19] 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. 2

[20] Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*, 2023. 2, 4

[21] Chunyuan Li, Zhe Gan, Zhengyuan Yang, Jianwei Yang, Linjie Li, Lijuan Wang, and Jianfeng Gao. Multimodal foundation models: From specialists to general-purpose assistants. *arXiv preprint arXiv:2309.10020*, 2023. 2

[22] Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *arXiv preprint arXiv:2306.00890*, 2023. 2

[23] 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. 2, 4, 6

[24] Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. *arXiv preprint arXiv:2101.00190*, 2021. 3

[25] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*, 2023. 2, 4

[26] 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. 2

[27] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023. 2, 4, 5, 6

[28] 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. 2, 4, 6

[29] Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, 2022. 4

[30] Yadong Lu, Chunyuan Li, Haotian Liu, Jianwei Yang, Jianfeng Gao, and Yelong Shen. An empirical study of scaling instruct-tuned large multimodal models. *arXiv preprint arXiv:2309.09958*, 2023. 2, 4

[31] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 11–20, 2016. 4, 6

[32] Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 4, 6

[33] Anand Mishra, Shashank Shekhar, Ajeet Kumar Singh, and Anirban Chakraborty. Ocr-vqa: Visual question answering by reading text in images. In *2019 international conference on document analysis and recognition (ICDAR)*, pages 947–952. IEEE, 2019. 4, 6

[34] OpenAI. Gpt-4v(ision) system card. https://cdn.openai.com/papers/GPTV_System_Card.pdf, 2023. 2

[35] 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. 4

[36] Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. A-okvqa: A benchmark for visual question answering using world knowledge. In *European Conference on Computer Vision*, pages 146–162. Springer, 2022. 4, 6

[37] ShareGPT. https://sharegpt.com/, 2023. 4, 6

[38] Oleksii Sidorov, Ronghang Hu, Marcus Rohrbach, and Amanpreet Singh. Textcaps: a dataset for image captioning with reading comprehension. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 742–758. Springer, 2020. 4, 6

[39] 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. 4

[40] 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. 2, 3, 4

[41] Shilong Zhang, Peize Sun, Shoufa Chen, Min Xiao, Wenqi Shao, Wenwei Zhang, Kai Chen, and Ping Luo. Gpt4roi: Instruction tuning large language model on region-of-interest. *arXiv preprint arXiv:2307.03601*, 2023. 2

[42] Yanzhe Zhang, Ruiyi Zhang, Jiuxiang Gu, Yufan Zhou, Nedim Lipka, Diyi Yang, and Tong Sun. Llavar: Enhanced visual instruction tuning for text-rich image understanding. *arXiv preprint arXiv:2306.17107*, 2023. 2

[43] Bo Zhao, Boya Wu, and Tiejun Huang. Svit: Scaling up visual instruction tuning. *arXiv preprint arXiv:2307.04087*, 2023. 2

[44] Yunqing Zhao, Tianyu Pang, Chao Du, Xiao Yang, Chongxuan Li, Ngai-Man Cheung, and Min Lin. On evaluating adversarial robustness of large vision-language models. *arXiv preprint arXiv:2305.16934*, 2023. 2

[45] Chunting Zhou, Pengfei Liu, Puxin Xu, Srini Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, et al. Lima: Less is more for alignment. *arXiv preprint arXiv:2305.11206*, 2023. 2

[46] 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. 2