

To See is to Believe: Prompting GPT-4V for Better Visual Instruction Tuning

Junke Wang^{1*} Lingchen Meng^{1*} Zejia Weng¹ Bo He² Zuxuan Wu^{1†} Yu-Gang Jiang¹

¹Fudan University ² University of Maryland

Abstract

Existing visual instruction tuning methods typically prompt large language models with textual descriptions to generate instruction-following data. Despite the promising performance achieved, these descriptions are derived from image annotations, which are oftentimes coarse-grained. Furthermore, the instructions might even contradict the visual content without observing the entire visual context. To address this challenge, we introduce a fine-grained visual instruction dataset, LVIS-INSTRUCT4V, which contains 220K visually aligned and context-aware instructions produced by prompting the powerful GPT-4V with images from LVIS. Through experimental validation and case studies, we demonstrate that high-quality visual instructional data could improve the performance of LLaVA-1.5, a state-of-the-art large multimodal model, across a wide spectrum of benchmarks by clear margins. Notably, by simply replacing the LLaVA-Instruct with our LVIS-INSTRUCT4V, we achieve better results than LLaVA on most challenging LMM benchmarks, *e.g.*, LLaVA^w (76.7 vs. 70.7) and MM-Vet (40.2 vs. 35.4). We release our data and model at <https://github.com/X2FD/LVIS-INSTRUCT4V>.

1 Introduction

Recent years have witnessed the booming development of Large Language Models (LLM) [5, 50, 40, 41], which represent a remarkable advancement in AI by offering unprecedented language understanding and reasoning capabilities. The great potential of LLMs motivates researchers in the computer vision community to build large multimodal models (LMM) on top of LLMs such that visual inputs can be effectively analyzed.

Visual instruction tuning [25] offers a promising way to bridge visual and textual information. Exemplary approaches, such as LLaVA [22, 24] and MiniGPT-4 [51], have demonstrated exceptional proficiency in natural instruction-following and visual reasoning. Typically, these methods leverage large language models [29, 30] to produce a large amount of instruction-answer pairs based on the textual description of the images, *e.g.*, captions or bounding boxes with labels. In the absence of direct access to visual signals, crucial details are sometimes missing, and in some cases, the instructions are even incorrect, as depicted in Figure 1.

To address this issue, we propose to harness GPT-4V, a powerful large multimodal model, to curate more accurate, context-aware, and fine-grained instruction-following data with a carefully designed pipeline. In particular, we feed GPT-4V with images from LVIS [13], an object detection dataset characterized by a comprehensive taxonomy and intricate annotations, along with their corresponding box annotations, and prompt the model to generate two types of instruction-answer pairs: contextually-

* Equal contributions.

† Corresponding author.

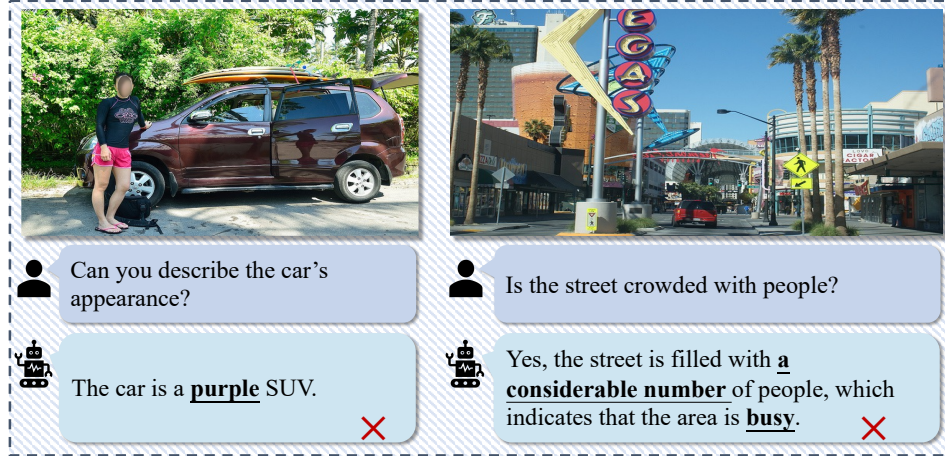


Figure 1: Examples of incorrect instruction-answer pairs in LLaVA [22], which are generated by prompting GPT-4. Without direct access to the images, GPT-4 may produce incorrect responses when asked fine-grained questions about color, location, *etc.*

aware conversations related to the visual contents and high-quality image descriptions. Finally, we have 220K detailed instructions, forming a large-scale instruction dataset, LVIS-INSTRUCT4V.

We adopt the same model architecture with LLaVA-1.5 [24], and replace their LLaVA-Instruct with our LVIS-INSTRUCT4V for visual instruction tuning. The experimental results demonstrate that significant improvements could be observed in most traditional QA benchmarks [12, 14] and more challenging LMM benchmarks, *e.g.*, LLaVA^w (76.7 vs. 70.7) and MM-Vet (40.2 vs. 35.4). In addition, the overall results could be further improved if we mix our data with LLaVA-Instruct (see Figure 2 for a better comparison with other state-of-the-art methods).

2 Related Work

2.1 Large Multimodal Models

Large Multimodal Models (LMMs) [1, 8, 21, 23, 10], also known as Multimodal Large Language Models (MLLMs), have attracted emerging attention due to their profound impact for a wide range of applications like image captioning [47, 46] and visual question-answering [2, 38]. These methods are typically built upon Large Language Models (LLMs) to extend their powerful comprehension and reasoning capability to the multimodal domain. Frozen [42] and Flamingo [1] represent the pioneering attempts of LMMs, which are trained on interleaved visual and textual data for in-context few-shot learning. However, they directly feed visual tokens into language models without an intermediary projector, which presents a remarkable challenge for cross-modal alignment. To mitigate this issue, BLIP-2 [23] proposes to connect the image encoder to a pretrained LLM through a lightweight Q-former. With this, two-stage pretraining is conducted to enforce the Q-Former to learn the textual grounded and LLM-interpretable visual representations, respectively.

2.2 Visual Instruction Tuning

Instruction tuning [32, 44, 45] is first explored in the field of NLP, which finetunes or adaptes language models [5, 34, 50, 40] to follow specific textual instructions. Despite the prominent success in the NLP domain, the inherent text-only nature poses limitations on its ability to effectively comprehend and interact with visual content. To address this challenge, LLaVA [22] introduces *visual instruction tuning* for the first time, enabling language models to follow visual instructions and engage with multimodal information. After that, a series of approaches have been developed to improve the visual instruction tuning with advanced architectures [16, 9, 4] or more versatile functionalities [33, 51, 20]. Nevertheless, the instruction data employed by these methods are generated either by manually reformatting the existing annotated datasets like VQA [12], or prompting GPT-4 [30] to derive from the textual description of images, which undoubtedly limits the fine-grained visual understanding of

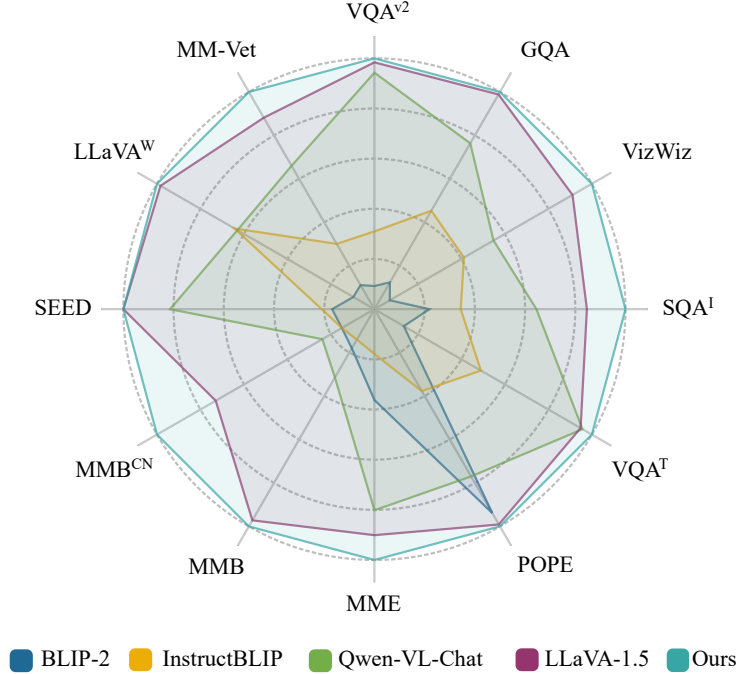


Figure 2: Comparisons of with other SOTA methods on various benchmarks. Please refer to Table 1 for detailed numbers.

the results models. In this paper, we make use of the strong multimodal understanding capability of GPT-4V to collect a novel visual instruction dataset, LVIS-INSTRUCT4V. Benefiting from the direct access to images during the data generation process, the instruction-answer pairs in LVIS-INSTRUCT4V are not only more visually aligned but also encompass rich visual details.

3 LVIS-INSTRUCT4V

Data collection. The impressive success of Large Language Models (LLMs) [23, 1] has highlighted the potential of data scaling in training high-capacity models. However, the majority of existing multimodal datasets have predominantly been tailored for specific tasks, such as image-level understanding [6, 35] or region-level referring comprehension [48, 17]. While these datasets serve their respective purposes, they often lack the depth and complexity of data necessary to address more intricate challenges rooted in multi-turn interactions.

In contrast to prior task-specific designs, the acquisition of instruction-following data requires multi-turn interactions. Notably, LLaVA [25] uses ChatGPT/GPT-4 to autonomously generate instruction-following data by leveraging location and semantics from detection data. However, the instructions are generated by a language-only approach using available annotations. This limitation can lead to the omission of crucial details or introduce inaccuracies when visual inputs are absent.

In this paper, we aim to curate a fine-grained instruction-following dataset by explicitly conditioning the generation of instruction data on both language and visual inputs. This ensures that instructions are carefully crafted based on visual context, taking into account the categories and attributes of objects, as well as their interrelationships. To achieve this goal, we leverage image data from LVIS [13], as well as their fine-grained annotations of an extensive taxonomy¹, to prompt GPT-4V to undertake two key tasks: (1) generate conversational question-answer lists through self-reasoning and (2) produce high-quality image descriptions guided by precise bounding box information. For a comprehensive overview of the detailed prompts used in data generation, please refer to the Appendix A. To the end, we use 110K images from LVIS and generate 220K high-quality visual instructions. Figure 3 gives

¹Images in LVIS [13] are all sourced from COCO, but the annotations are different.

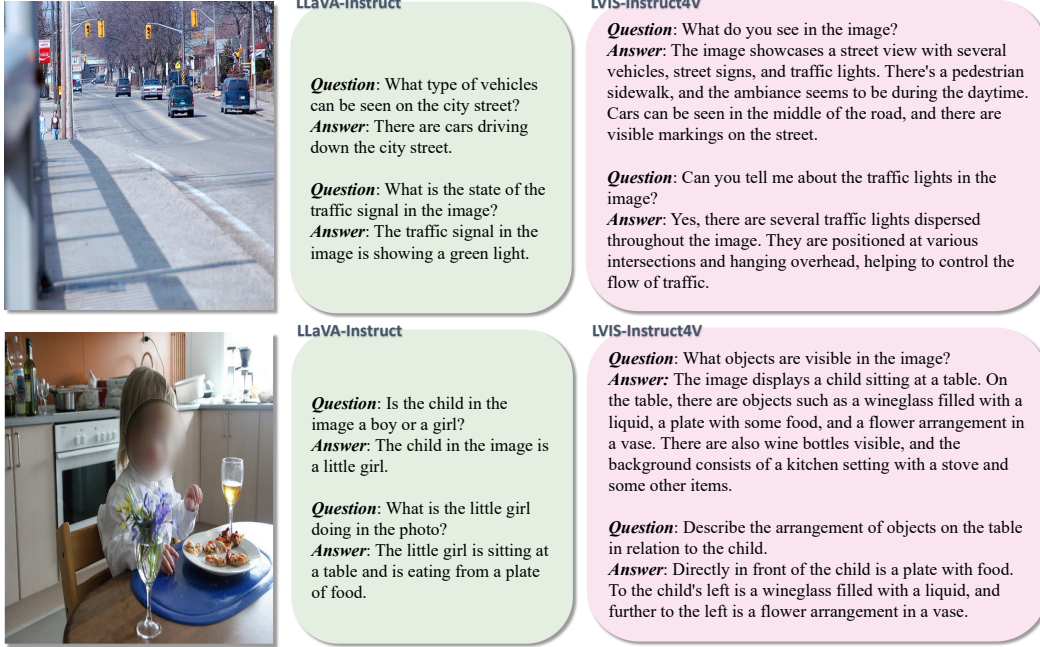


Figure 3: Examples of the instruction-answer pairs from LLaVA and our instruction dataset.

an example of generated instructions and comparisons with those from LLaVA. We can observe that instructions based on visual and language inputs are more detailed.

Instruction analysis. To gain a better understanding of instructions derived with and without visual inputs, we investigate the distribution of words related to “position”, “count”, “size”, “color”, “material”, and “shape”, which are fine-grained clues, for LLaVA-Instruct and LVIS-Instruct4V. We then instantiate each clue with a list of words. For example, words like “black”, “blue”, “brown”, “gray”, etc are used as indicators of color. Please refer to Appendix B for the complete list. We then calculate the occurrence of words within the instructions. The resulting counts are normalized by the total number of instructions. Figure 4 shows the results. We can observe that instructions from LVIS-INSTRUCT4V indeed exhibit a greater richness in terms of fine-grained information.

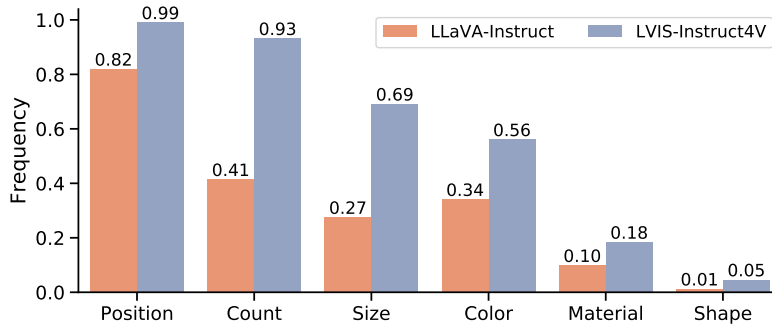


Figure 4: Statistics of instruction data of LLaVA-Instruct and LVIS-Instruct4V.

4 Experiment

4.1 Implementation Details

We adopt the same model architectures with LLaVA-1.5 [24]. By default, our models are trained using a combination of LVIS-INSTRUCT4V and benchmark data, a collection of task-related data [12, 27, 15, 28, 36, 39, 18, 19, 37], following [24]. We also additionally incorporate LLaVA-Instruct for

Table 1: Comparison with state-of-the-art LMMs on various benchmarks. w/ G: use GPT models to generate instruction-following data, Priv: the data are private. * denotes the training images of the datasets are observed during training. V-7/13B indicates Vicuna-7/13B [43]; L-7/13B indicates LLaMA-7/13B [40, 41]. Q-7B indicates Qwen-7B [3]. The best results are marked in bold.

(a) Results on traditional VQA benchmarks.

Method	LLM	w/ G	VQA ^{v2}	GQA	VisWiz	SQA ^I	VQA ^T
InstructBLIP [9]	V-7B	GPT-4	–	49.2	34.5	60.5	50.1
IDEFICS-9B [16]	L-7B	✗	50.9	38.4	35.5	–	25.9
Qwen-VL [4]	Q-7B	Priv	78.8*	59.3*	35.2	67.1	63.8
Qwen-VL-chat [4]	Q-7B	Priv	78.2*	57.5*	38.9	68.2	61.5
LLaVA-1.5 [24]	V-7B	GPT-4	78.5*	62.0*	50.0	66.8	58.2
Ours	V-7B	GPT-4V	79.2*	62.6*	52.5	70.3	57.6
Ours-mixLLaVA	V-7B	GPT-4 & 4V	79.6*	62.6*	51.8	70.1	58.7
BLIP-2 [23]	V-13B	✗	41.0	41.0	19.6	61.0	42.5
InstructBLIP [9]	V-13B	GPT-4	–	49.5	33.4	63.1	50.7
IDEFICS-80B [16]	L-65B	✗	60.0	45.2	36.0	–	30.9
Shikra [7]	V-13B	GPT-4	77.4*	–	–	–	–
LLaVA-1.5 [24]	V-13B	GPT-4	80.0*	63.3*	53.6	71.6	61.3
Ours	V-13B	GPT-4V	80.1*	63.8*	51.4	73.0	62.1
Ours-mixLLaVA	V-13B	GPT-4 & 4V	80.7*	63.6*	57.2	74.2	62.5

(b) Results on recent LMM benchmarks.

Method	LLM	w/ G	POPE	MME	MMB	MMB ^{CN}	SEED	LLaVA ^W	MM-Vet
InstructBLIP [9]	V-7B	GPT-4	–	–	36.0	23.7	53.4	60.9	26.2
IDEFICS-9B [16]	L-7B	✗	–	–	48.2	25.2	–	–	–
Qwen-VL [4]	Q-7B	Priv	–	–	38.2	7.4	56.3	–	–
Qwen-VL-chat [4]	Q-7B	Priv	–	1487.5	60.6	56.7	58.2	–	–
LLaVA-1.5 [24]	V-7B	GPT-4	85.9	1510.7	64.3	58.3	58.6	63.4	30.5
Ours	V-7B	GPT-4V	84.0	1472.9	67.1	67.1	60.8	70.4	34.6
Ours-mixLLaVA	V-7B	GPT-4 & 4V	86.0	1528.2	66.2	66.4	60.6	67.0	31.5
BLIP-2 [23]	V-13B	✗	85.3	1293.8	–	–	46.4	38.1	22.4
InstructBLIP [9]	V-13B	GPT-4	78.9	1212.8	–	–	–	58.2	25.6
IDEFICS-80B [16]	L-65B	✗	–	–	54.5	38.1	–	–	–
Shikra [7]	V-13B	GPT-4	–	–	58.8	–	–	–	–
LLaVA-1.5 [24]	V-13B	GPT-4	85.9	1531.3	67.7	63.6	61.6	70.7	35.4
Ours	V-13B	GPT-4V	85.3	1572.0	67.8	68.2	62.5	76.7	40.2
Ours-mixLLaVA	V-13B	GPT-4 & 4V	86.0	1574.9	68.0	67.4	61.6	71.3	37.4

data scaling experiments. The batch size, image resolution, and training schedules are also the same as LLaVA-1.5.

4.2 Comparison with State-of-the-art Methods

In Table 1, we present a comprehensive comparison between the model trained using our data, and other large multimodal models (LMMs) across a diverse array of benchmarks. Using Vicuna-7B as the language model, our method achieves 79.2 on VQAv2 [12], and 67.1 on MMBench-CN [26]. When the size of the language model is scaled to 13B, the results on them are further improved to 80.1 and 68.2, respectively. Notably, our 7B and 13B models significantly outperform the counterparts of LLaVA-1.5 models across various recently proposed challenging LMM benchmarks, *e.g.*, LLaVA^w and MM-Vet, by only changing the LLaVA-Instruct to our LVIS-INSTRUCT4V.

In order to further boost the performance, we additionally use LLaVA-Instruct during training, and the resulting model achieves state-of-the-art results on the majority of these benchmarks, specifically, 11 out of 12. Remarkably, employing the same architecture and training procedure, it surpasses

LLaVA [24] across all benchmarks. We achieve a substantial 43.6 gain (1574.9 vs. 1531, measured by scores) on MME [11] and an impressive 3.6% improvement on MM-Vet [49]. It is worth pointing out that MME and MM-Vet are challenging benchmarks designed to evaluate LMMs, which contain fine-grained multimodal tasks like counting, OCR, *etc.* These results highlight the effectiveness of LVIS-INSTRUCT4V.

4.3 Discussion



Ablation studies. We study the effects of optimization and training schedules on instruction tuning in Table 2. The results in rows 1 and 2 show that even if the language model is kept frozen during instruction finetuning (Ours-13B-mixLLaVA-fzLLM), our data could still improve the performance on SQA, MME, and SEED compared with LLaVA-1.5, validating its superior quality. In addition, the performance on MME could be boosted to 1581.3 if we finetune the LLaVA-1.5 model on our data (Ours-13B-mixLLaVA-ftLLaVA-1.5), surpassing its original score by 50.0 (1531.3 vs. 1581.3). Finally, we also find that training the model for a longer schedule (Ours-13B-mixLLaVA-2epoch) could not bring additional performance gains, and training for one epoch as LLaVA-1.5 leads to overall better results.

Table 2: Ablation studies on both traditional VQA benchmarks and LMM benchmarks.

Method	VQA ^{v2}	VisWiz	SQA ¹	MME	SEED
LLaVA-1.5-13B	80.0	53.6	71.6	1531.3	61.6
Ours-13B-mixLLaVA-fzLLM	80.0	50.0	74.2	1544.3	61.7
Ours-13B-mixLLaVA-ftLLaVA-1.5	80.6	51.9	72.7	1581.3	61.6
Ours-13B-mixLLaVA-2epoch	80.5	56.3	72.7	1538.4	61.6
Ours-13B-mixLLaVA	80.7	57.2	74.2	1574.9	61.6

Visualizations. We now provide qualitative visualization of results produced by different models in Table 3. We can see that by additionally training with LVIS-INSTRUCT4V, our model is better at recognizing small texts and identifying the relationships among objects.

Table 3: Our model achieves more accurate scene text recognition and finer-grained differentiation.

Scene text recognition:		Fine-grained visual differentiation:	
			
User	What are the scene texts in the image?	User	There are two images (left and right). Now you have two captions: CaptionA: “the blue bowl is placed on the right of the small red coffee mug”, and CaptionB: “the white bowl is on the right of the small red coffee mug”. Please indicate which caption corresponds to the left image and which caption corresponds to the right image.
LLaVA-1.5	The scene text in the image reads “Capilano Suspension Bridge.”	LLaVA-1.5	CaptionA corresponds to the right image and CaptionB corresponds to the left image.
Ours	The scene texts in the image are “Capilano Suspension Bridge” and “ Be astonished by the natural beauty. ”	Ours	CaptionA corresponds to the left image and CaptionB corresponds to the right image.

5 Conclusion

Existing instructions for training large multimodal models are typically generated by prompting language models with textual descriptions, which are derived from image annotations. While offering decent results, without ever seeing images, the performance of these instructions is upper-bounded by manual annotations in current benchmarks, which exist in the forms of captions and bounding boxes. The reasons are two-fold: (1) these annotations are often times coarse-grained; (2) LLMs can generate inaccurate answers without the entire visual context. To address these challenges, we introduced a large-scale visual instruction dataset LVIS-INSTRUCT4V, which contains 220K fine-grained and visually aligned instructions. LVIS-INSTRUCT4V is curated by harnessing the great visual analysis ability of GPT-4V through carefully designed prompts. We then built upon LLaVA, a state-of-the-art of multimodal language model, to evaluate the efficacy of LVIS-INSTRUCT4V. We conducted extensive experiments on a wide range of benchmarks and the results clearly demonstrate that the proposed dataset benefits visual instruction tuning.

References

- [1] J.-B. Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc, A. Mensch, K. Millican, M. Reynolds, et al. Flamingo: a visual language model for few-shot learning. *arXiv preprint arXiv:2204.14198*, 2022.
- [2] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh. Vqa: Visual question answering. In *ICCV*, 2015.
- [3] J. Bai, S. Bai, Y. Chu, Z. Cui, K. Dang, X. Deng, Y. Fan, W. Ge, Y. Han, F. Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.
- [4] J. Bai, S. Bai, S. Yang, S. Wang, S. Tan, P. Wang, J. Lin, C. Zhou, and J. Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 2023.
- [5] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al. Language models are few-shot learners. In *NeurIPS*, 2020.
- [6] S. Changpinyo, P. Sharma, N. Ding, and R. Soricut. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *CVPR*, 2021.
- [7] K. Chen, Z. Zhang, W. Zeng, R. Zhang, F. Zhu, and R. Zhao. Shikra: Unleashing multimodal llm’s referential dialogue magic. *arXiv preprint arXiv:2306.15195*, 2023.
- [8] A. Chowdhery, S. Narang, J. Devlin, M. Bosma, G. Mishra, A. Roberts, P. Barham, H. W. Chung, C. Sutton, S. Gehrmann, et al. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*, 2022.
- [9] W. Dai, J. Li, D. Li, A. M. H. Tiong, J. Zhao, W. Wang, B. Li, P. Fung, and S. Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *arXiv preprint arXiv:2305.06500*, 2023.
- [10] D. Driess, F. Xia, M. S. Sajjadi, C. Lynch, A. Chowdhery, B. Ichter, A. Wahid, J. Tompson, Q. Vuong, T. Yu, et al. PaLM-E: An embodied multimodal language model. *arXiv preprint arXiv:2303.03378*, 2023.
- [11] C. Fu, P. Chen, Y. Shen, Y. Qin, M. Zhang, X. Lin, Z. Qiu, W. Lin, J. Yang, X. Zheng, et al. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*, 2023.
- [12] Y. Goyal, T. Khot, D. Summers-Stay, D. Batra, and D. Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *CVPR*, 2017.
- [13] A. Gupta, P. Dollar, and R. Girshick. Lvis: A dataset for large vocabulary instance segmentation. In *CVPR*, 2019.

- [14] D. Gurari, Q. Li, A. J. Stangl, A. Guo, C. Lin, K. Grauman, J. Luo, and J. P. Bigham. Vizwiz grand challenge: Answering visual questions from blind people. In *CVPR*, 2018.
- [15] D. A. Hudson and C. D. Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *CVPR*, 2019.
- [16] IDEFICS. Introducing idefics: An open reproduction of state-of-the-art visual language model. <https://huggingface.co/blog/idefics>, 2023.
- [17] A. Kamath, M. Singh, Y. LeCun, G. Synnaeve, I. Misra, and N. Carion. Mdetr-modulated detection for end-to-end multi-modal understanding. In *CVPR*, 2021.
- [18] S. Kazemzadeh, V. Ordonez, M. Matten, and T. Berg. Referitgame: Referring to objects in photographs of natural scenes. In *EMNLP*, 2014.
- [19] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *IJCV*, 2017.
- [20] X. Lai, Z. Tian, Y. Chen, Y. Li, Y. Yuan, S. Liu, and J. Jia. Lisa: Reasoning segmentation via large language model. *arXiv preprint arXiv:2308.00692*, 2023.
- [21] C. Li, Z. Gan, Z. Yang, J. Yang, L. Li, L. Wang, and J. Gao. Multimodal foundation models: From specialists to general-purpose assistants. *arXiv preprint arXiv:2309.10020*, 2023.
- [22] C. Li, C. Wong, S. Zhang, N. Usuyama, H. Liu, J. Yang, T. Naumann, H. Poon, and J. Gao. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *arXiv preprint arXiv:2306.00890*, 2023.
- [23] J. Li, D. Li, S. Savarese, and S. Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023.
- [24] H. Liu, C. Li, Y. Li, and Y. J. Lee. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*, 2023.
- [25] H. Liu, C. Li, Q. Wu, and Y. J. Lee. Visual instruction tuning. In *NeurIPS*, 2023.
- [26] Y. Liu, H. Duan, Y. Zhang, B. Li, S. Zhang, W. Zhao, Y. Yuan, J. Wang, C. He, Z. Liu, et al. Mmbench: Is your multi-modal model an all-around player? *arXiv preprint arXiv:2307.06281*, 2023.
- [27] K. Marino, M. Rastegari, A. Farhadi, and R. Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *CVPR*, 2019.
- [28] A. Mishra, S. Shekhar, A. K. Singh, and A. Chakraborty. Ocr-vqa: Visual question answering by reading text in images. In *ICDAR*, 2019.
- [29] OpenAI. Chatgpt. <https://openai.com/blog/chatgpt/>, 2023.
- [30] OpenAI. Gpt-4 technical report, 2023.
- [31] OpenAI. Gpt-4v(ision) system card. https://cdn.openai.com/papers/GPTV_System_Card.pdf, 2023.
- [32] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, et al. Training language models to follow instructions with human feedback. In *NeurIPS*, 2022.
- [33] Z. Peng, W. Wang, L. Dong, Y. Hao, S. Huang, S. Ma, and F. Wei. Kosmos-2: Grounding multimodal large language models to the world. *arXiv preprint arXiv:2306.14824*, 2023.
- [34] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *JMLR*, 2020.

- [35] C. Schuhmann, R. Beaumont, R. Vencu, C. Gordon, R. Wightman, M. Cherti, T. Coombes, A. Katta, C. Mullis, M. Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. In *NeurIPS*, 2022.
- [36] D. Schwenk, A. Khandelwal, C. Clark, K. Marino, and R. Mottaghi. A-okvqa: A benchmark for visual question answering using world knowledge. In *ECCV*, 2022.
- [37] ShareGPT. <https://sharegpt.com/>, 2023.
- [38] K. J. Shih, S. Singh, and D. Hoiem. Where to look: Focus regions for visual question answering. In *CVPR*, 2016.
- [39] A. Singh, V. Natarajan, M. Shah, Y. Jiang, X. Chen, D. Batra, D. Parikh, and M. Rohrbach. Towards vqa models that can read. In *CVPR*, 2019.
- [40] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.
- [41] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.
- [42] M. Tsimpoukelli, J. L. Menick, S. Cabi, S. Eslami, O. Vinyals, and F. Hill. Multimodal few-shot learning with frozen language models. In *NeurIPS*, 2021.
- [43] Vicuna. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. <https://vicuna.lmsys.org/>, 2023.
- [44] Y. Wang, Y. Kordi, S. Mishra, A. Liu, N. A. Smith, D. Khashabi, and H. Hajishirzi. Self-instruct: Aligning language model with self generated instructions. *arXiv preprint arXiv:2212.10560*, 2022.
- [45] Y. Wang, S. Mishra, P. Alipoormolabashi, Y. Kordi, A. Mirzaei, A. Arunkumar, A. Ashok, A. S. Dhanasekaran, A. Naik, D. Stap, et al. Benchmarking generalization via in-context instructions on 1,600+ language tasks. *arXiv preprint arXiv:2204.07705*, 2022.
- [46] T. Yao, Y. Pan, Y. Li, Z. Qiu, and T. Mei. Boosting image captioning with attributes. In *ICCV*, 2017.
- [47] Q. You, H. Jin, Z. Wang, C. Fang, and J. Luo. Image captioning with semantic attention. In *CVPR*, 2016.
- [48] L. Yu, P. Poirson, S. Yang, A. C. Berg, and T. L. Berg. Modeling context in referring expressions. In *ECCV*, 2016.
- [49] W. Yu, Z. Yang, L. Li, J. Wang, K. Lin, Z. Liu, X. Wang, and L. Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*, 2023.
- [50] S. Zhang, S. Roller, N. Goyal, M. Artetxe, M. Chen, S. Chen, C. Dewan, M. Diab, X. Li, X. V. Lin, et al. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022.
- [51] D. Zhu, J. Chen, X. Shen, X. Li, and M. Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023.

A Pseudo code for instruction data generation

Algorithm 1: Pseudo code for instruction data generation with GPT-4V [31].

```
1 PROMPT_DICT{
2   prompt_conversation: (
3     "You are an AI visual assistant, and you are seeing a single image. Answer all questions as you
      are seeing the image. Design a conversation between you and a person asking about this photo.
      The answers should be in a tone that a visual AI assistant is seeing the image and answering the
      question. Ask diverse questions and give corresponding answers. Include questions asking
      about the visual content of the image, including the object types, counting the objects, object
      actions, object locations, relative positions between objects, etc. Only include questions that
      have definite answers: (1) one can see the content in the image that the question asks about and
      can answer confidently; (2) one can determine confidently from the image that it is not in the
      image. Do not ask any questions that cannot be answered confidently. Also include complex
      questions that are relevant to the content in the image, for example, asking about background
      knowledge of the objects in the image, asking to discuss events happening in the image, etc.
      Again, do not ask about uncertain details, but the questions should be challenging enough,
      requiring the person to utilize 1) complex reasoning; 2) world knowledge; 3) explanatory
      answers; and 4) multi-turn conversation, to give accurate answers. Please provide detailed
      answers when answering complex questions. For example, give detailed examples or reasoning
      steps to make the content more convincing and well-organized. Please ensure all the questions
      are closely related to the visual content of the provided image, which means that if the person
      cannot see the picture but only gets access to the text description of it, he/she will not be able to
      answer accurately. If the AI assistant asks counterfactual questions, the person should give a
      negative answer, rather than making up an answer."
4   ),
5   prompt_detail: (
6     "You are an AI visual assistant that can analyze a single image. You will receive several sets of
      descriptions of the objects in the image, each in the format: set id, category of the 1st object:
      location of the 1st object; category of the 2nd object: location of the 2nd object; .... Note that
      different sets are separated by a link break. These locations are in the form of bounding boxes,
      represented as (x1, y1, x2, y2) with floating numbers ranging from 0 to 1. These values
      correspond to the top left x, top left y, bottom right x, and bottom right y. Utilize the supplied
      bounding box details to formulate a precise and comprehensive caption that accurately describes
      the highlighted objects within their designated areas, regardless of the objects'sequence. Note
      that please describe the overall information of all objects in each set, rather than giving each
      object in a set an independent description. To generate high-quality captions, you should first
      understand the visual content, then based on the background knowledge or reasoning, either
      explain why the things are happening that way or provide guidance to facilitate the better
      understanding of these objects. Ensure that the caption is tightly connected to the objects in
      question, meticulously avoiding any reference to extraneous objects beyond the defined
      locations. If addressing the query accurately becomes too challenging, please respond with
      [failed]. Instead of directly mentioning the bounding box coordinates, utilize this data to explain
      the scene using natural language. Include details like object counts, the position of the objects,
      relative position between the objects. When using the information from the categories and
      coordinates, directly explain the scene, and do not mention that the information source is the
      caption or the bounding box. Always answer as if you are directly looking at the image. Please
      proceed with the tasks in the set order, respectively. Make sure that the caption for each set
      begins with "Bbox List", followed by a set id. The object information about categories and
      normalized coordinates are as follows: Category, (x1, y1, x2, y2); Category,
      (x1, y1, x2, y2); ..."
7   )
8 }
9 output = openai.ChatCompletion.create(
10     model="gpt-4v",
11     messages=["role": "user", "content": Image; prompt ],
12 )
```

B Vocabulary List

We define a list of vocabularies for instruction analysis.

Type	Words
Position	'above', 'ahead', 'ahead position', 'back', 'back direction', 'backward', 'behind', 'behind position', 'below', 'down', 'downward', 'downward direction', 'elevated', 'elevation', 'forward', 'forward direction', 'front', 'front position', 'frontward', 'higher', 'left', 'left direction', 'leftward', 'lower', 'lower position', 'more ahead', 'more back', 'more behind', 'more forward', 'more front', 'more left', 'more right', 'on', 'on position', 'overhead', 'overhead position', 'right', 'right direction', 'rightward', 'underneath', 'underneath position', 'up', 'upward', 'upward direction'
Count	'one', 'two', 'three', 'four', 'five', 'six', 'seven', 'eight', 'nine', 'ten'
Size	'average', 'averageness', 'big', 'bigness', 'considerably', 'enormity', 'enormous', 'giant', 'giantness', 'huge', 'hugeness', 'large', 'largeness', 'larger', 'little', 'littleness', 'massive', 'massiveness', 'medium', 'medium size', 'middling', 'mini', 'miniature', 'minuteness', 'moderate', 'moderately', 'moderateness', 'more average', 'more enormous', 'more middling', 'more miniature', 'more moderate', 'petite', 'petiteness', 'slightly', 'small', 'smaller', 'smallness', 'tininess', 'tiny'
Color	'black', 'blue', 'brown', 'gray', 'green', 'orange', 'pink', 'purple', 'red', 'white', 'yellow'
Material	'cotton', 'glass', 'glassy', 'iron', 'leather', 'linen', 'linen-like', 'metal', 'plastic', 'polymeric', 'satiny', 'silk', 'steel', 'stone', 'stony', 'timber', 'velvet', 'wooden', 'woody'
Shape	'circle', 'circular', 'hexagon', 'hexagonal', 'octagon', 'octagonal', 'oval', 'rectangle', 'rectangular', 'square', 'triangle', 'triangular'

Table 4: List of words for instruction analysis.