The All-Seeing Project V2: Towards General Relation Comprehension of the Open World

Weiyun Wang[2,1],Yiming Ren[3,1],Haowen Luo[3],Tiantong Li[3,1],Chenxiang Yan[3], Zhe Chen[5,1], Wenhai Wang[4,1], Qingyun Li[6,1], Lewei Lu[7], Xizhou Zhu[3,1,7], Yu Qiao[1], and Jifeng Dai[†3,1]

[1]OpenGVLab, Shanghai AI Laboratory [2]Fudan University
[3]Tsinghua University [4]The Chinese University of Hong Kong
[5]Nanjing University [6]Harbin Institute of Technology [7]SenseTime Research

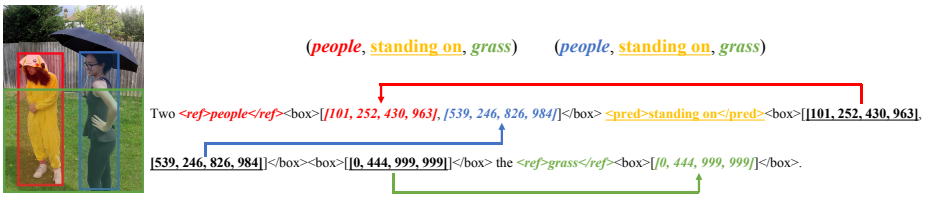


Fig. 1: A complex example for the formulation of Relation Conversation.

A Relation Conversation

In this section, we introduce more details about the formulation of Relation Conversation. As depicted in Fig. 1, when a certain text span is associated with multiple regions, these bounding boxes are formatted as:

$$\texttt{<box>[[}x_1^1, y_1^1, x_2^1, y_2^1\texttt{], ..., [}x_1^n, y_1^n, x_2^n, y_2^n\texttt{]]</box>},$$

where $\texttt{[}x_1^i, y_1^i, x_2^i, y_2^i\texttt{]}$ denotes the i-th bounding box linked to the object or predicate. For a specific predicate, the subject and object must be linked to an equal number of bounding boxes. Otherwise, one of them must be linked to just one bounding box and thus can be broadcast to match the count of another one.

As shown in Fig. 1, to parse this example into a scene graph, we first assign the semantic tag "`people`" to the bounding box highlighted in red and blue. Similarly, we assign the semantic tag "`grass`" to the bounding box highlighted in green. We then extract the predicate label enclosed in "`<pred></pred>`" (*i.e.*, "`standing on`") and the box coordinates of its subjects and objects (*i.e.*, bounding boxes highlighted with **bold underline**). After that, we utilize these box coordinates as keys to match their respective semantic tags. Considering that two subjects are linked to the predicate while only one object is linked, we broadcast the object to match the number of subjects. Given N subjects and N objects, we pack them into N tuples in order, where each tuple consists of one subject and one object. In this example, we obtain two tuples, resulting in two parsed triplet `(people, standing on, grass)`, each connected to a `people` with different bounding boxes.

B The All-Seeing Dataset v2

More data examples of AS-V2 are shown in Figs. 2 to 4. Besides, prompts used to generate detailed description data are shown in Tabs. 1 and 2.

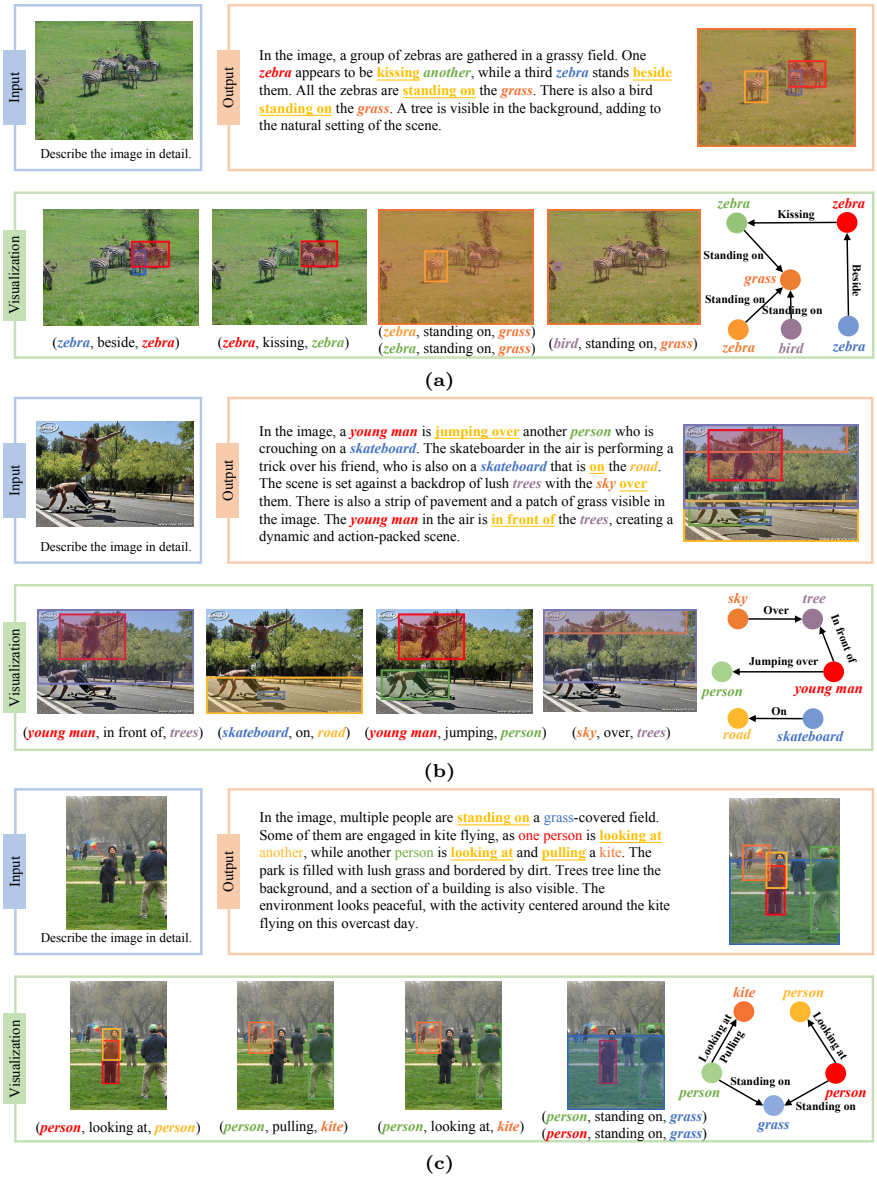


Fig. 2: Data Examples of Detailed Description task in AS-V2.

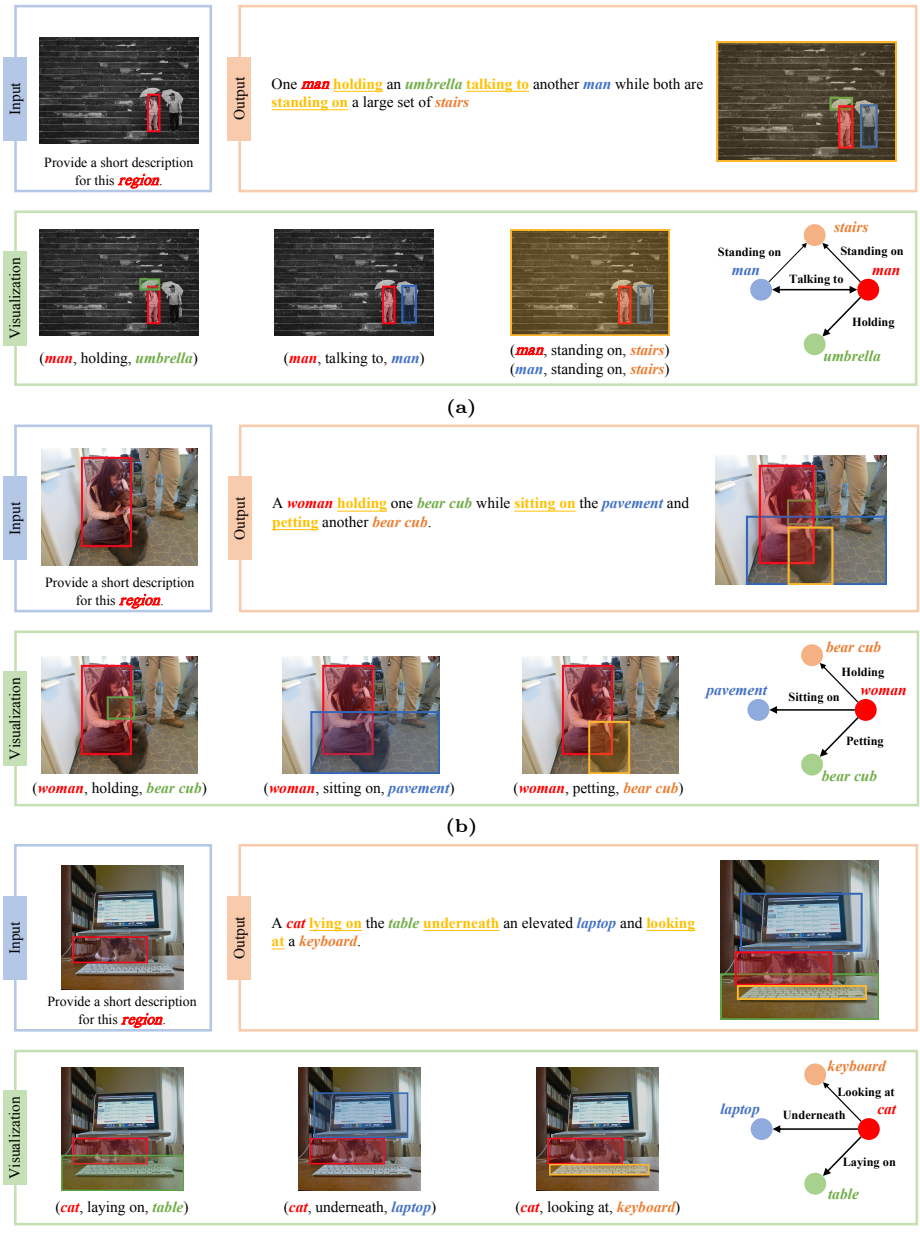


Fig. 3: Data Examples of Region Captioning task in AS-V2.

Fig. 4: Data Examples of Conversation task in AS-V2. Due the space limitations, we exhibit only one turn for each conversation.

Table 1: For each query, the system info explains the **task description** and the in-context-learning examples are presented in the form of multi-turn conversation. For each turn, the input `query['context']` consists of (1) the **image** to be annotated, (2) the **captions annotations** of this image, (3) the **location annotations**, as well as (4) the **relation annotations**. The output `query['context']` comprises the manually annotated scene graph conversation data. In this example, we provide the task description for the Detailed Description data in the relation conversation.

```
messages = [{"role":"system", "content": f"""You are an AI visual assistant that can analyze a single image. You receive one image and five sentences, each describing this image you are observing. In addition, specific object locations within the image are given, along with detailed coordinates. These coordinates are in the form of bounding boxes, represented as [x1, y1, x2, y2] with int numbers ranging from 0 to 999. These values correspond to the top left x, top left y, bottom right x, and bottom right y. Note that these coordinates are normalized. Besides, the scene graph of this image is also provided as a list of tuples. Each tuple is represented as (subject, bounding box of the subject, object, bounding box of the object, predicate).

Using the provided caption, bounding box, and scene graph information, describe the scene in a detailed manner. If there are errors in the caption, please ignore them and do not point them out in your description.

Instead of directly mentioning the bounding box coordinates, utilize this data to explain the scene using natural language with its bounding box in the format like "<ref>object</ref><box>[[x1, y1, x2, y2]]</box>". When mentioning the predicate between two objects, you should mention it in the format like "<pred>predicate</pred><box>[[x1, y1, x2, y2]]</box><box>[[x3, y3, x4, y4]]</box>", where "<box>[[x1, y1, x2, y2]]</box>" denotes the bounding box coordinates of the subject and "<box>[[x3, y3, x4, y4]]</box>" denotes the bounding box coordinates of the object. Include details like object counts, position of the objects, relative position between the objects.

When using the information from the caption, coordinates, or scene graph, directly explain the scene, and do not mention that the information source is the caption or the bounding box or the scene graph. You should mention all tuples and predicates included in the scene graph in the generated caption. Make sure that the box following the <pred>predicate</pred> has already been mentioned after a <ref>object</ref>."""}
]
for sample in fewshot_samples:
    messages.append({"role":"user","content":sample['context']})
    messages.append({"role":"assistant","content":sample['response']})
messages.append({"role":"user","content":'\n'.join(query)})
```

Table 2: One example to illustrate the instruction-following data. The top block shows the contexts such as captions, locations, relations and images used to prompt GPT, and the bottom block shows the three types of responses. Note that the visual image is also used to prompt GPT.

Context type 1: Captions
A pretty young lady holding a dark colored umbrella.
A girl in a Pikachu suit standing beside a girl holding an umbrella.
A woman is wearing a raincoat and another woman is holding an umbrella over her head.
two girls stand in a yard, one holds an umbrella and the other is dressed as pikachu.
Two women who are standing in the grass.

Context type 2: Locations
person: [101, 252, 430, 963]
person: [539, 246, 826, 984]
grass: [0, 444, 999, 999]
...

Context type 3: Relations
(person, [101, 252, 430, 963], grass, [0, 444, 999, 999], standing on)
(person, [101, 252, 430, 963], person, [539, 246, 826, 984], beside)
(person, [539, 246, 826, 984], grass, [0, 444, 999, 999], standing on)
...

Response type 1: Detailed Description
In the image, two <ref>people</ref><box>[[101, 252, 430, 963], [539, 246, 826, 984]]</box> are <pred>standing on</pred><box>[[101, 252, 430, 963], [539, 246, 826, 984]]</box><box>[[0, 444, 999, 999]]</box> a <ref>grass</ref><box>[[0, 444, 999, 999]]</box>. ...

Response type 2: Region Captioning
person: [539, 246, 826, 984]
A <ref>girl</ref><box>[[101, 252, 430, 963]]</box> in a Pikachu suit standing <pred>beside</pred><box>[[101, 252, 430, 963]]</box><box>-[[539, 246, 826, 984]]</box> a <ref>girl</ref><box>[[539, 246, 826, 984]]-</box>.
...

Response type 3: Conversation
Question:
What are the two <ref>people</ref><box>[[101, 252, 430, 963], [539, 246, 826, 984]]</box> doing in the image?
—
Answer:
In the image, one <ref>person</ref><box>[[539, 246, 826, 984]]</box> is <pred>holding</pred><box>[[539, 246, 826, 984]]</box><box>[[326, 117, 976, 435]]</box> an <ref>umbrella</ref><box>[[326, 117, 976, 435]]</box>, ...
...

Table 3: Details of the instruction-tuning data for ASMv2 in stage 2. We collect a wide range of high-quality data, totaling approximately 4 million samples.

Task	#Samples	Dataset
Captioning	124K	TextCaps [21], ShareGPT4V [3]
VQA	314K	VQAv2 [6], GQA [7], OKVQA [17], A-OKVQA [19], ScienceQA [15], CLEVR [8], Visual7W [31]
OCR	157K	ST-VQA [1], LLaVAR [29], OCR-VQA [18], DocVQA [4]
Grounding	643K	RefCOCO/+/g [9, 16]
RegionVQA	2.3M	RefCOCOg [16], VG [10], VCR [27], AS-Core [23]
Conversation	500K	LLaVA-Instruct [13], SVIT [30], LRV [12], AS-V2 (ours)
Text	40K	ShareGPT [20]

C The All-Seeing Model v2

C.1 Implementation Details

Training Stage 1. The global batch size is set to 256 in the pre-training phase and 128 in the instruction-tuning phase. We employ the AdamW optimizer [14] with the β_1 of 0.9, the β_2 of 0.999, and the weight decay of 0. The learning rate is initialized as 1×10^{-3} for the pre-training phase and 2×10^{-5} for the instruction-tuning phase. Both phases include a linear warmup that lasts until the first 3% of training steps. The warmup is followed by a cosine decay strategy with a minimum learning rate of 0. We only train the vision-language connector in the pre-training phase while both the vision-language connector and the language model are trainable in the instruction-tuning phase. We train the model for 1 epoch in both phases. The image resolution of ASMv2 is set to 336 $\times$ 336.

Training Stage 2. The global batch size is set to 512 and the learning rate is initialized as 2×10^{-5} in both the pre-training phase and the instruction-tuning phase. The language model and vision-language connector are trainable in both phases while the vision encoder is always frozen. We train the model for 5000 steps in the pre-training phase and 1 epoch in the instruction-tuning phase. The other settings remain the same as the instruction-tuning phase of Stage 1.

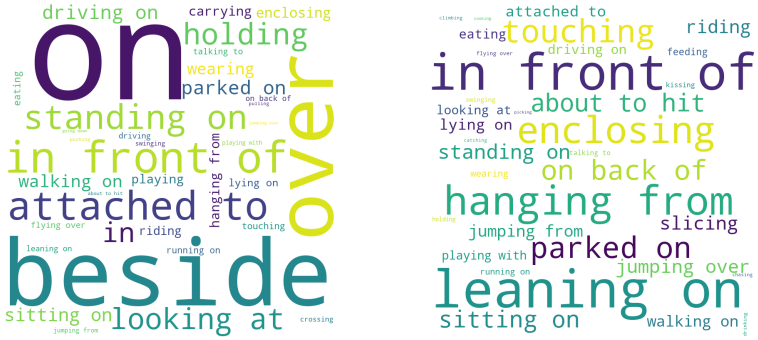
C.2 Predicate Classification

In this section, we evaluate the relation comprehension capability of our model through the Predicate Classification task (PredCls) on the Panoptic Scene Graph (PSG) dataset [25]. Compared to the Open-ended Scene Graph Generation task, PredCls aims to generate a scene graph given the ground-truth object labels and localization, focusing on the relation prediction performance without the interference of the detection performance.

Evaluation Setup. Assuming that the number of ground-truth objects is N, we query the model for $N \times (N-1)$ times, considering each of the ground-truth

Table 4: Recall scores on Predicate Classification task.

Method	Predicate Classification					
	R@20	mR@20	R@50	mR@50	R@100	mR@100
IMP [24]	30.5	9.0	35.9	10.5	38.3	11.3
MOTIFS [28]	45.1	19.9	50.5	21.5	52.5	22.2
VCTree [22]	**45.9**	**21.4**	**51.2**	23.1	**53.1**	23.8
GPSNet [11]	38.8	17.1	46.6	20.2	50.0	21.3
ASMv2 (ours)	17.6	**21.4**	25.9	**34.4**	32.6	**44.5**



(a) (b)

Fig. 5: Word Clouds for evaluation data in PSG. Fig. 5a visualizes the distribution of ground-truth predicates while Fig. 5b visualizes those predicted by ASMv2.

objects as the subject or object. For each query, we ask the model "*What is the relation between the* `<subject>` *and the* `<object>`*? Answer the question using a single word or phrase.*" and employ a vocabulary ranking method [2] to generate the scores for each predicate label. Following prior works [22, 25], we report the Recall and mean Recall (mRecall) here.

Results. As shown in Tab. 4, our ASMv2 demonstrates competitive performance on the Predicate Classification task within the PSG dataset. Specifically, our ASMv2 achieves superior performance in mRecall but is inferior in Recall. For instance, our ASMv2 significantly outperforms VCTree by 11.3 points in mR@50 and 20.7 points in mR@100, while it falls behind in terms of Recall. These results stem from the PSG dataset's inherently imbalanced distribution of predicate labels, where broad predicates such as "on" and "over" are more frequent. As depicted in Fig. 5, our ASMv2 is less likely to predict these common but general predicates. Instead, it tends to predict more specific and less frequent predicates, like "standing on" and "parked on", resulting in superior mRecall while inferior Recall. These results underline our model's deeper and more detailed comprehension of visual relations.

Table 5: Ablation results. We report the $Q \rightarrow AR$ accuracy for VCR [27] and the overall accuracy for CRPE. REC denotes the average accuracy score across RefCOCO [9], RefCOCO+ [16], and RefCOCOg [16].

Ablation Settings	MME	MM-Vet	REC	VCR	CRPE
ASMv2	1621.0	41.3	87.4	78.4	64.5
- Stage1 Training	1527.6	35.6	86.1	78.7	64.7
- Relation Conversation	1554.6	42.2	86.6	77.0	55.3

C.3 Ablation Study

In this section, we ablate the training settings of ASMv2. The experimental settings are the same as those discussed in Appendix C.1. As shown in Tab. 5, the two-stage training process of ASMv2 and the utilization of relation conversation data is essential for achieving excellent performance on both image-level and region-level benchmarks simultaneously. We can observe that skipping the first training stage leads to significant performance degradation on MME [5], MM-Vet [26], and Referring Expression Comprehension (REC) Benchmarks [9, 16], indicating that the model struggles to understand visual information at both the image and region levels simultaneously without the two-stage training strategy.

Furthermore, removing relation conversation data from the training corpora results in inferior performance on REC [9, 16], VCR [27], and CRPE. The performance on MME [5] also experiences a drop of about 66.4 points, primarily due to a decrease in the count score, from 170.0 to 155.0, and a decrease in the position score, from 163.3 to 133.3. These results demonstrate the effectiveness of our relation conversation data in improving capabilities for region-level visual information understanding and relation comprehension.

D The Circular-based Relationship Probing Evaluation

In this section, we present more examples of abnormal data in CRPE in Fig. 6.

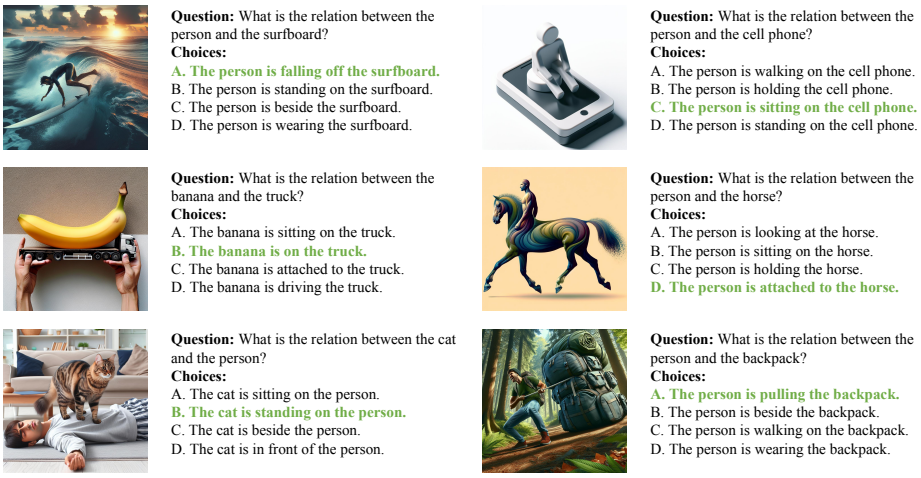


Fig. 6: Data examples of abnormal data in the CRPE.

References

1. Biten, A.F., Tito, R., Mafla, A., Gomez, L., Rusinol, M., Valveny, E., Jawahar, C., Karatzas, D.: Scene text visual question answering. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4291–4301 (2019)
2. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners. NeurIPS (2020)
3. Chen, L., Li, J., Dong, X., Zhang, P., He, C., Wang, J., Zhao, F., Lin, D.: Sharegpt4v: Improving large multi-modal models with better captions. arXiv preprint arXiv:2311.12793 (2023)
4. Clark, C., Gardner, M.: Simple and effective multi-paragraph reading comprehension. arXiv preprint arXiv:1710.10723 (2017)
5. Fu, C., Chen, P., Shen, Y., Qin, Y., Zhang, M., Lin, X., Qiu, Z., Lin, W., Yang, J., Zheng, X., et al.: Mme: A comprehensive evaluation benchmark for multimodal large language models. arXiv preprint arXiv:2306.13394 (2023)
6. Goyal, Y., Khot, T., Summers-Stay, D., Batra, D., Parikh, D.: Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In: CVPR (2017)
7. Hudson, D.A., Manning, C.D.: Gqa: A new dataset for real-world visual reasoning and compositional question answering. In: CVPR (2019)

8. Johnson, J., Hariharan, B., Van Der Maaten, L., Fei-Fei, L., Lawrence Zitnick, C., Girshick, R.: Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In: CVPR (2017)
9. Kazemzadeh, S., Ordonez, V., Matten, M., Berg, T.: Referitgame: Referring to objects in photographs of natural scenes. In: EMNLP (2014)
10. Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.J., Shamma, D.A., et al.: Visual genome: Connecting language and vision using crowdsourced dense image annotations. International journal of computer vision (2017)
11. Lin, X., Ding, C., Zeng, J., Tao, D.: Gps-net: Graph property sensing network for scene graph generation. In: CVPR. pp. 3746–3753 (2020)
12. Liu, F., Lin, K., Li, L., Wang, J., Yacoob, Y., Wang, L.: Aligning large multi-modal model with robust instruction tuning. arXiv preprint arXiv:2306.14565 (2023)
13. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. arXiv preprint arXiv:2304.08485 (2023)
14. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: ICLR (2019)
15. Lu, P., Mishra, S., Xia, T., Qiu, L., Chang, K.W., Zhu, S.C., Tafjord, O., Clark, P., Kalyan, A.: Learn to explain: Multimodal reasoning via thought chains for science question answering. NeurIPS (2022)
16. Mao, J., Huang, J., Toshev, A., Camburu, O., Yuille, A.L., Murphy, K.: Generation and comprehension of unambiguous object descriptions. In: CVPR (2016)
17. Marino, K., Rastegari, M., Farhadi, A., Mottaghi, R.: Ok-vqa: A visual question answering benchmark requiring external knowledge. In: CVPR (2019)
18. Mishra, A., Shekhar, S., Singh, A.K., Chakraborty, A.: Ocr-vqa: Visual question answering by reading text in images. In: ICDAR. pp. 947–952. IEEE (2019)
19. Schwenk, D., Khandelwal, A., Clark, C., Marino, K., Mottaghi, R.: A-okvqa: A benchmark for visual question answering using world knowledge. In: ECCV (2022)
20. ShareGPT: https://sharegpt.com/ (2023)
21. Sidorov, O., Hu, R., Rohrbach, M., Singh, A.: Textcaps: a dataset for image captioning with reading comprehension. In: ECCV. pp. 742–758. Springer (2020)
22. Tang, K., Zhang, H., Wu, B., Luo, W., Liu, W.: Learning to compose dynamic tree structures for visual contexts. In: CVPR. pp. 6619–6628 (2019)
23. Wang, W., Shi, M., Li, Q., Wang, W., Huang, Z., Xing, L., Chen, Z., Li, H., Zhu, X., Cao, Z., et al.: The all-seeing project: Towards panoptic visual recognition and understanding of the open world. arXiv preprint arXiv:2308.01907 (2023)
24. Xu, D., Zhu, Y., Choy, C.B., Fei-Fei, L.: Scene graph generation by iterative message passing. In: CVPR (2017)
25. Yang, J., Ang, Y.Z., Guo, Z., Zhou, K., Zhang, W., Liu, Z.: Panoptic scene graph generation. In: ECCV (2022)
26. Yu, W., Yang, Z., Li, L., Wang, J., Lin, K., Liu, Z., Wang, X., Wang, L.: Mm-vet: Evaluating large multimodal models for integrated capabilities. arXiv preprint arXiv:2308.02490 (2023)
27. Zellers, R., Bisk, Y., Farhadi, A., Choi, Y.: From recognition to cognition: Visual commonsense reasoning. In: CVPR (2019)
28. Zellers, R., Yatskar, M., Thomson, S., Choi, Y.: Neural motifs: Scene graph parsing with global context. In: CVPR (2018)
29. Zhang, Y., Zhang, R., Gu, J., Zhou, Y., Lipka, N., Yang, D., Sun, T.: Llavar: Enhanced visual instruction tuning for text-rich image understanding. arXiv preprint arXiv:2306.17107 (2023)
30. Zhao, B., Wu, B., Huang, T.: Svit: Scaling up visual instruction tuning. arXiv preprint arXiv:2307.04087 (2023)

31. Zhu, Y., Groth, O., Bernstein, M., Fei-Fei, L.: Visual7w: Grounded question answering in images. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4995–5004 (2016)