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

Abstract. We present the All-Seeing Project V2: a new model and dataset designed for understanding object relations in images. Specifically, we propose the All-Seeing Model V2 (ASMv2) that integrates the formulation of text generation, object localization, and relation comprehension into a relation conversation (ReC) task. Leveraging this unified task, our model excels not only in perceiving and recognizing all objects within the image but also in grasping the intricate relation graph between them, diminishing the relation hallucination often encountered by Multi-modal Large Language Models (MLLMs). To facilitate training and evaluation of MLLMs in relation understanding, we created the first high-quality ReC dataset (AS-V2) which is aligned with the format of standard instruction tuning data. In addition, we design a new benchmark, termed Circular-based Relation Probing Evaluation (CRPE) for comprehensively evaluating the relation comprehension capabilities of MLLMs. Notably, our ASMv2 achieves an overall accuracy of 64.50 on this relation-aware benchmark, surpassing the 55.63 of LLaVA-1.5 by a large margin. We hope that our work can inspire more future research and contribute to the evolution towards artificial general intelligence. Our project is released at https://github.com/OpenGVLab/all-seeing.

Keywords: Multimodal Large Language Model · Pointer instructions

1 Introduction

The study of artificial general intelligence (AGI) systems that can match human intelligence and excel in any task across domains represents the ultimate goal in the field of artificial intelligence. Benefiting from the advancements of Large Language Models (LLMs), Multi-modal Large Language Models (MLLMs) have demonstrated impressive capabilities in a variety of Vision-Language tasks, suggesting new avenues for achieving AGI. However, as shown in Fig. 1a, most popular MLLMs [11,50,51] are limited to understanding images as a whole.

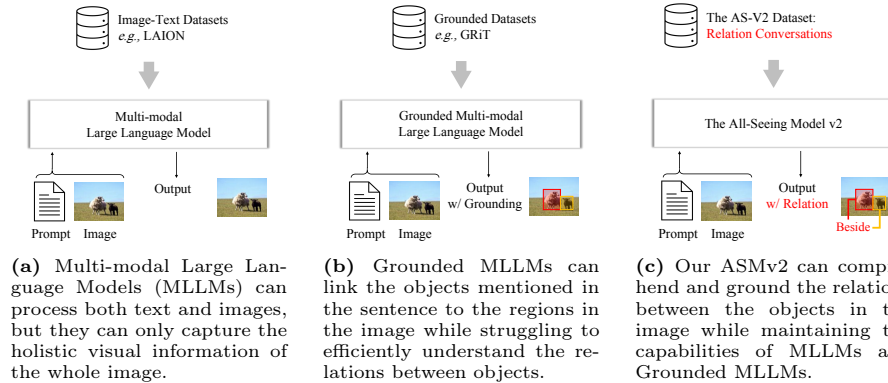[†]Corresponding to Jifeng Dai <daijifeng@tsinghua.edu.cn>.

(a) Multi-modal Large Language Models (MLLMs) can process both text and images, but they can only capture the holistic visual information of the whole image.

(b) Grounded MLLMs can link the objects mentioned in the sentence to the regions in the image while struggling to efficiently understand the relations between objects.

(c) Our ASMv2 can comprehend and ground the relations between the objects in the image while maintaining the capabilities of MLLMs and Grounded MLLMs.

Fig. 1: Overview and comparison of our All-Seeing Model v2 with other MLLMs. To enhance relation comprehension ability while maintaining grounding, referring, and other general capabilities, we propose (1) a novel task, termed Relation Conversation (ReC), which unifies the formulation of text generation, object localization, and relation comprehension; (2) a high-quality dataset AS-V2, which consists of over 127K samples for ReC; (3) the All-Seeing Model v2 (ASMv2), which is capable of comprehending and grounding the relations between the objects in the image.

As an effective method to improve interaction efficiency, the capabilities of grounding and referring (*i.e.*, adopting bounding boxes in responses) have attracted increasing attention and have been widely integrated into current Grounded MLLMs [2,8,63,81,82]. Such capabilities empower models to provide visual responses (*e.g.*, bounding boxes), supporting more vision-language tasks such as region captioning [32,58], referring expression comprehension [31,58], and referring question answering [92]. However, as shown in Fig. 1b, existing models primarily focus on recognizing certain objects within images, overlooking the perception of relations between these objects. Due to the lack of appropriate modeling methods and suitable training data for relation knowledge, these models struggle to comprehend the inter-object relations within images accurately. Consequently, these models are prone to hallucinations when dealing with relation questions or overly relying on language priors for judgment.

To enhance relation comprehension ability while maintaining grounding, referring, and other general capabilities, we introduce a novel task, termed Relation Conversation (ReC). The formulation of ReC unifies the modeling of text generation, object localization, and relation comprehension. Specifically, as depicted in Fig. 1c, ReC requires the model to generate the text response while linking all mentioned objects as well as the subjects and objects of each predicate in the response to the corresponding regions in the image simultaneously. Such explicit requirement for predicate grounding challenges the model to comprehend relations between objects within the image. Notably, models trained on ReC can be naturally adapted to the Scene Graph Generation task. The grounded objects serve as the nodes in the scene graph while the grounded predicates serve as the edges. Compared with the traditional scene graph generation, ReC enables the model to generate the scene graph in an open-ended manner, demonstrat-
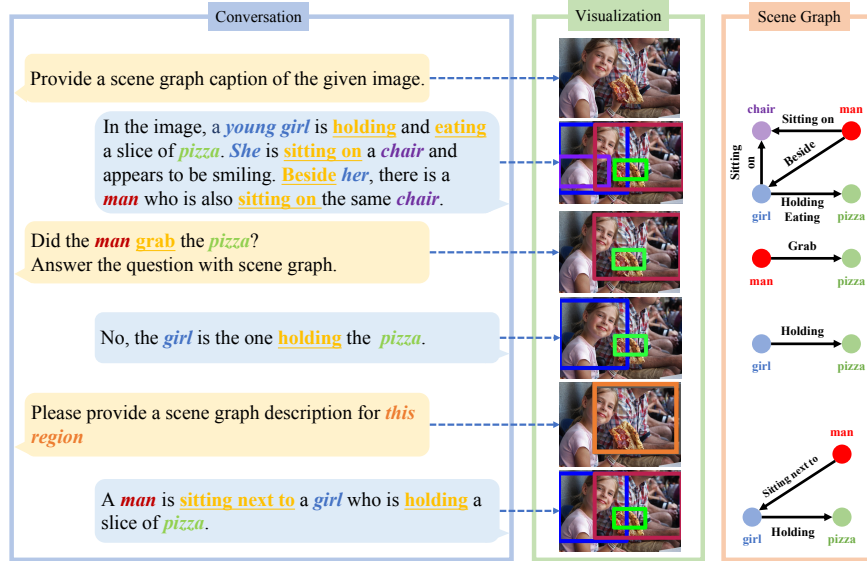
Fig. 2: Examples of relation conversation responses from ASMv2.

ing the potential to generalize to previously unseen predicate labels, while also maintaining the general ability of MLLMs.

From the data aspect, we construct the All-Seeing Dataset V2 (AS-V2) comprising 127K high-quality relation conversation samples, which is built upon the existing caption [10], location [46], and relation [85] annotations. Combining AS-V2 with other image-level and region-level multimodal corpora for training, we propose the All-Seeing Model v2 (ASMv2). Benefiting from the tailored task format and data, our model can deal with three types of relation tasks, including (1) Relation Conversation, which requires the model to link all mentioned objects and predicates to the corresponding regions in the image; (2) Open-ended Scene Graph Generation, which requires the model to generate a scene graph based on the given image in an open-ended manner; (3) Predicate Classification, which requires the model to generate a scene graph given the ground-truth object labels and localization. An example of ASMv2 is shown in Fig. 2.

To evaluate the relation comprehension ability of existing MLLMs, we construct a benchmark called Circular-based Relation Probing Evaluation (CRPE), which is the first benchmark that covers all elements of the relation triplets `(subject, predicate, object)`, providing a systematic platform for the evaluation of relation comprehension ability. CRPE is formulated as single-choice questions and consists of four splits: Existence, Subject, Predicate, and Object. The Existence split evaluates the object recognition ability while the remaining splits are designed to evaluate the relation comprehension capability. Additionally, to evaluate the dependency on language priors, we include abnormal data in CRPE, which depict relations that are rare but reasonable in the real world.

Our main contributions are as follows:

(1) We introduce the All-Seeing Project V2, which endows MLLMs with the ability not only to perceive all objects within the image but also to recognize the relations between these objects, leading to superior relation comprehension capability and the potential to generate scene graphs in an open-ended manner.

(2) We propose a novel task, termed Relation Conversation, and the corresponding formulation method, unifying the modeling of captioning, grounding, and relation tasks flexibly. Based on the task and formulation, we constructed the AS-V2 dataset. Combining AS-V2 with other general multimodal corpora for training, we propose the All-Seeing Model v2 (ASMv2), which demonstrates powerful performance across various tasks, including Open-ended Scene Graph Generation and other general image-level and region-level vision-language tasks.

(3) To evaluate the relation comprehension ability of existing MLLMs, we construct the CRPE benchmark. Notably, our ASMv2 achieves an overall accuracy of 64.50 on CRPE, surpassing the 55.63 of LLaVA-1.5 by a large margin. We also evaluate ASMv2 on various image-level and region-level vision-language tasks. Specifically, our model achieves an overall score of 74.4 on MMBench [52] and 1621.0 on MME [18], surpassing LLaVA-1.5 [50] by 5.5 points and 90.0 points separately. Besides, the average accuracy of ASMv2 on grounding benchmarks [31, 58] is 87.42, outperforming Qwen-VL [2] by 1.69 points.

2   Related Work

2.1   Vision-Language Models

Significant advancements have been made in the field of visual recognition and understanding in recent years. Models based on the image-text matching framework [11, 17, 29, 66] achieve powerful zero-shot performance on various downstream tasks, thereby initiating the era of open-world semantic recognition and understanding. Subsequent works [38, 88] further integrate this framework with language modeling tasks to support more generative tasks. The recent progress of Large Language Models [5, 61, 78] leads to the emergency of many LLM-based multimodal models [11, 36, 39, 51, 77, 82, 90, 95, 107], which aim to integrate the powerful understanding and reasoning ability of LLMs with multimodal perception and comprehension. Despite their powerful performance, these works are only capable of capturing the holistic visual information of the whole image. Some recent methods [8, 33, 53, 62, 63, 67, 81, 96, 99] begin to focus on location-aware understanding. However, due to the lack of appropriate modeling methods and training data for relation comprehension, these methods struggle to comprehend the inter-object relations within images accurately. To enhance relation comprehension ability while maintaining other general capabilities of MLLMs, we introduce a novel task, termed Relation Conversation, which unifies the formulation of text generation, object localization, and relation comprehension.

2.2   Scene Graph Generation

Scene Graph Generation (SGG) [55] is a crucial task in scene understanding and has attracted substantial interest across the research community. This area

has witnessed the proposal of diverse model architectures, including message-passing-based frameworks [14, 22, 24, 42, 43, 93], attention-based networks [65, 104], tree-structured networks [26, 97], and DETR-based networks [13, 40, 71]. While most existing methods only utilize images as input, recent works begin to incorporate language information or knowledge graphs to facilitate SGG [16, 27, 44, 55, 98, 106], although the scope of language utilization remains limited to basic object or relation concepts. Compared to prior specialized models, our model is a powerful general model with strong vision-language understanding and reasoning ability and can generate the scene graph in an open-ended manner, exhibiting the potential to generalize to previously unseen predicate labels.

2.3   Benchmarks for Relation Comprehension

Evaluating the comprehension of relations between objects is a crucial aspect of advancing MLLMs. Benchmarks like Visual Genome [32] and COCO [10, 46] provide foundational datasets for object detection and image captioning. These datasets primarily focus on individual object recognition and general descriptive capabilities. They include annotations for object relations but are not explicitly designed to probe the depth of relation comprehension in a structured and focused manner. Some synthetic datasets [1, 30, 75], are introduced to probe the spatial reasoning capabilities of vision-language models. These datasets offer controlled environments for model evaluation but inherently limit the problem's scope due to their bounded expressivity. The Visual Spatial Reasoning (VSR) dataset [48] asks the model to classify whether the caption correctly describes the relation of two objects presented in the image. This approach primarily focuses on binary classification tasks instead of the understanding of relations within the scene. In this work, we introduce the CRPE benchmark, which consists of different splits and each split is designed to probe one of the elements in the relation triplet (`subject, predicate, object`). Therefore, we can evaluate the relation comprehension ability of existing MLLMs more systematically.

3   Data Construction

3.1   The All-Seeing Dataset v2

Our objective is to establish a dataset to unlock the Relation Conversation capability for Multi-modal Large Language Models (MLLMs), which requires the model to predict not only the bounding boxes of each object but also those of the subjects and objects for each predicate mentioned in the sentence. In this section, we elaborate on the method for constructing the training dataset for ReC, termed All-Seeing Dataset v2 (AS-V2). Specifically, we utilize GPT-4V [62] to construct AS-V2 based on COCO images [6] and their annotations [10, 46, 85]. The key idea is to query GPT-4V to generate responses while linking the objects and predicates mentioned in the generated response to specific regions within the image, referring to the given location annotations and relation annotations. The formulation of Relation Conversation is presented in Sec. 4.1.
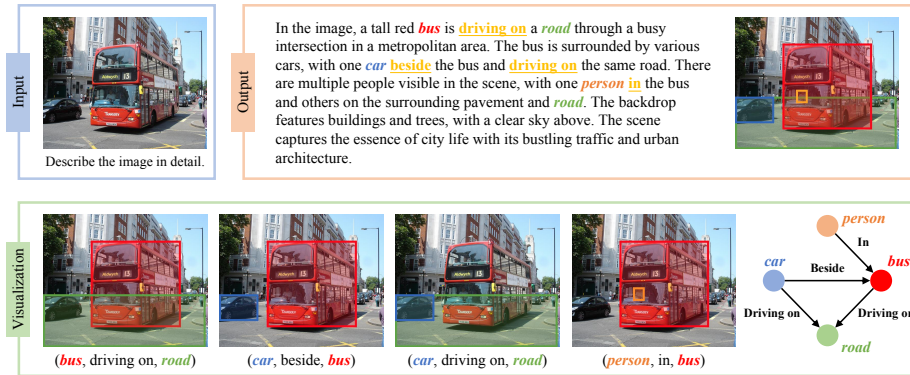
Fig. 3: Data example in the AS-V2 dataset. In the relation conversation, all mentioned objects are linked to their corresponding regions within the image while the predicates are linked to the regions corresponding to their subjects and objects.

The prompt for GPT-4V comprises six components: (1) **Task description**, which explains the formulation of relation conversation. (2) **Image** to be annotated. (3) **Caption annotations** of this image, intended to enhance GPT-4V's understanding of the scene. (4) **Location annotations**, which are the bounding boxes of the objects in the scene and guide GPT-4V in annotating the objects in the caption. (5) **Relation annotations**, which are presented as a list of (subject, predicate, object) triplets and help GPT-4V to annotate the predicate in the caption. (6) **Seed examples**, which are manually annotated to assist GPT-4V in comprehending the task description and formatting the output. Although the caption annotations are not necessary for GPT-4V to produce the desired relation conversation data, incorporating these details into the prompt significantly reduces the hallucinations in the generated data. An example of the prompt is presented in Appendix B.

The generated data comprise three types, including: (1) **Detailed description**, which requires the model to generate a comprehensive description for an image. (2) **Region captioning**, which requires the model to generate a comprehensive description for a certain region within the image. (3) **Conversation**, which requires the model to respond to the user query in the multi-turn conversation. The question types include the relations between objects, the object types, counting the objects, object actions, object locations, and relative positions between objects. Each type of data is generated using different task descriptions and human-annotated seed examples. These tasks require the model to understand pointer instructions (*e.g.*, utilizing bounding boxes as prompts) and link the objects and predicates mentioned in the generated response to the image regions. An example is shown in Fig. 3, with additional examples in Appendix B.

In this way, we collected 127K relation conversation samples in total, including 42K in detailed descriptions, 63K in region captioning, and 22K in conversations (90K turns in total), respectively. The conversation samples also include negative instructions to enhance model robustness. These instances contain incorrect relations, and the model should be able to recognize their incorrectness.
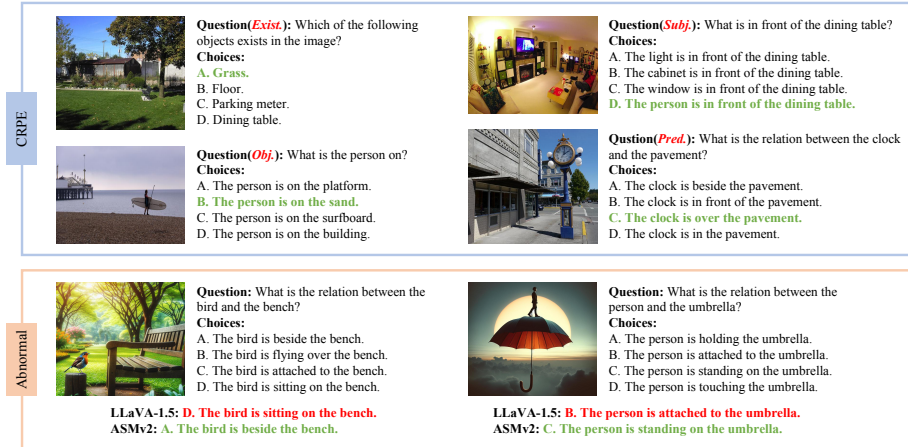
Fig. 4: Data examples in the CRPE. The benchmark is formulated as single-choice questions and consists of four splits: **Existence**, **Subject**, **Predicate**, and **Object**. The **Existence** split is designed to evaluate the object recognition ability while the remaining splits focus on the evaluation of relation comprehension. A qualitative comparison in abnormal data between LLaVA-1.5 and ASMv2 is shown at the bottom.

3.2 Circular-based Relation Probing Evaluation

In this section, we introduce CRPE, a benchmark designed to quantitatively evaluate the object recognition and relation comprehension capabilities of models. The evaluation is formulated as single-choice questions. For a robust evaluation, we adopt CircularEval [52] as our evaluation strategy. Under this setting, a question is considered as correctly answered only when the model consistently predicts the correct answer in each of the N iterations, with N corresponding to the number of choices. In each iteration, a circular shift is applied to both the choices and the answer to form a new query for the model.

As shown in Fig. 4, each sample in our benchmark consists of an image and a single-choice question with one correct answer and three wrong answers. The location annotations [46] and the triplets (`subject, predicate, object`) in the relation annotations [85] are utilized to generate the evaluation data. We construct four evaluation splits, including (1) the **Existence** split: the question of this split is "*Which of the following objects exists in the images?*". The correct answer is sampled from the semantic tags that exist in the image while the incorrect answer is sampled from those not exist in the image. (2) the **Subject** split: we generate the question based on the template "*What is `<predicate>` the `<object>`?*" and consider the subject in the triplet as the correct answer. The negative subjects are sampled from other semantic tags that exist in the image. (3) the **Predicate** split: we generate the question based on the template "*What is the relation between `<subject>` and `<object>`?*" and consider the predicate in the triplet as the correct answer. The negative predicates are randomly sampled. Only the predicates satisfying $P(p|s) > 0$ and $P(p|o) > 0$ can be sampled, where p, s, o refer to predicates, subjects, and objects, separately. (4) the **Object**

split: we generate the question based on the template "*What is the* `<subject>` `<predicate>`*?*" and consider the object in the triplet as the correct answer. The negative objects are sampled from other semantic tags that exist in the image. To avoid reference ambiguity, we ensure that the semantic tags of the subject and object in each triplet are distinct in the image. We also manually verify the generated samples and filter those with ambiguous questions or choices.

Additionally, to evaluate the dependency on language priors, we further include abnormal data in the Predicate split, which depict relations that are rare but reasonable in the real world. Specifically, we first select relation triplets with minimal $P(p|s, o)$ and then employ DALLE-3 [3] to generate corresponding images for these triplets. Considering that the generated images might not match the specified triplets exactly, we perform a manual filtering process for these triplet-image pairs to ensure data quality. After that, we generate the evaluation data using the method mentioned above. Fig. 4 shows abnormal examples at the bottom and more examples of these abnormal data are presented in Appendix D.

4   The All-Seeing Model v2

ASMv2 is a powerful Multi-modal Large Language Model (MLLM), which integrates the Relation Conversation (ReC) ability while maintaining powerful general capabilities. Specifically, it follows the model architecture of LLaVA-1.5 [50], comprising a vision encoder, a vision-language connector, and a language model. The Vicuna-13B [103] and CLIP-ViT-L-336px [20] are utilized as the language and vision backbone. This model can deal with three types of relation tasks, including (1) Relation Conversation, which requires the model to link all objects and predicates mentioned in the response to the corresponding regions in the image. (2) Open-ended Scene Graph Generation, which requires the model to generate a scene graph based on the given image in an open-ended manner; (3) Predicate Classification, which requires the model to generate a scene graph given the ground-truth object labels and localization; In addition, our model is also capable of multi-modality dialogue tasks such as Image Captioning, Visual Question Answering, and Multi-turn conversation. Since the ReC task requires the model to link the objects and predicates to the corresponding regions in the image, our ASMv2 is also endowed with grounding and referring capabilities and exhibits state-of-the-art performance on region-level tasks.

4.1   Relation Conversation

In this section, we elaborate on the formulation of ReC. Our objective is to propose a task that can enhance the relation comprehension ability while maintaining grounding, referring, and other general capabilities.

As depicted in Fig. 5, we represent the sentence in the relation conversation as a text sequence. Specifically, our relation conversation marks the object and predicate in the sentence using `<ref></ref>` and `<pred></pred>`, respectively. Each marked object is followed by a bounding box, indicating its localization. Similarly, each predicate is followed by two bounding boxes, which
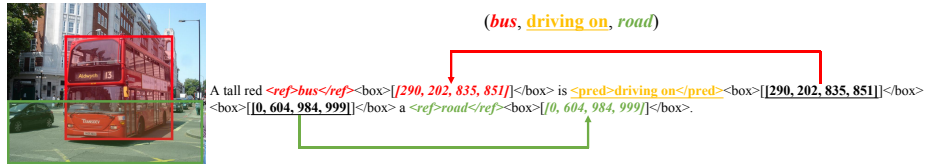
Fig. 5: Data formulation for Relation Conversation. Each object is marked with `<ref></ref>` and followed by a bounding box indicating its location while each predicate is marked with `<pred></pred>` and followed by two bounding boxes referring to its subjects and objects.

specifically refer to the subjects and objects of the predicate. All bounding boxes are normalized to integer values within the range [0, 1000) and formatted as: `<box>[[`x_1, y_1, x_2, y_2`]]</box>`. Please refer to Appendix A for more details.

Notably, the response in the relation conversation can be easily parsed into a scene graph. In a typical scene graph, each node denotes an object in the scene grounded by a bounding box with a semantic label, and each directed edge denotes the relation between a pair of objects with a predicate label. By utilizing the prediction of bounding boxes for each object (serving as semantic tags for nodes) and those for subjects and objects related to each predicate (serving as nodes, edges, and predicate labels), the generated ReC sentence can be naturally converted into a scene graph. Nodes without semantic tags will be labeled as Unknown. To convert the response shown in Fig. 5 into a scene graph, we first parse the objects marked by "`<ref></ref>`" and assign the marked text as the semantic tag of the following bounding box. Here, we assign "`bus`" and "`road`" as the semantic tag of the bounding box highlighted in red and green, separately. Then we extract the predicate label marked by "`<pred></pred>`" (*i.e.*, "`driving on`") and the box coordinates of the subjects and objects of it (*i.e.*, bounding boxes highlighted with **bold underline**). After that, we utilize these box coordinates as keys to match their semantic tags. In this example, the bounding box for the subject of "`driving on`" is matched with the box highlighted in red, therefore the subject of "`driving on`" is "`bus`". Similarly, the object of it is "`road`". Hence, we obtain the parsed triplet (`bus, driving on, road`). Note that the bounding boxes of the subject and object are also part of the triplet while we omit them for simplicity.

Compared with the traditional Scene Graph Generation task, our ReC task exhibits three advantages: (1) **More flexible.** Models trained on our proposed Relation Conversation task can be naturally adapted to the Scene Graph Generation task in an open-ended manner. (2) **Open-World.** Benefiting from the open-ended generation manner, models trained on ReC have the potential to generalize to previously unseen predicate labels in the Scene Graph Generation task. (3) **More general.** ReC requires models to generate a text response and link all mentioned objects to their corresponding regions within the image, thereby maintaining the grounding, referring, and general capabilities of MLLMs and broadening the applicability of these models in real-world scenarios.

4.2   Model Training

The training process of ASMv2 is divided into two stages, with each stage comprising a pre-training phase and an instruction-tuning phase. The first stage is designed to enable the model to effectively understand visual information at the image level. The pre-training phrase utilizes 595K samples from CC3M [70] filtered by LLaVA [51] while the instruction-tuning phrase utilizes a blend of 665K samples from LLaVA-1.5 [50]. We update the data format of the region-level data [31, 32, 58] in LLaVA-1.5 to the format introduced in Sec. 4.1. The second stage trains the model with a mixture of image-level data and region-level data, which enables the model to comprehend the visual information at the region level, facilitating effective grounding of objects and predicates within sentences. The pre-training phrase employs 5M samples from CC12M [70] filtered by BLIP [37], 10M filtered samples from AS-1B [81], and 15M filtered samples from GRiT [63]. The instruction-tuning phase employs 4M samples collected from a variety of sources, including image-level datasets [4, 9, 12, 30, 48–51, 57, 64, 100], region-level datasets [31, 32, 58, 81, 92, 101] and our proposed AS-V2 dataset. The summary of these datasets is presented in Tab. 9.

5   Experiments

In this section, we first compare our ASMv2 with leading Multi-modal Large Language Models (MLLMs) on representative vision-language benchmarks in Sec. 5.1. In addition to these image-level benchmarks, we also evaluate ASMv2 on three representative region-level tasks in Sec. 5.2. After that, ASMv2 is evaluated on the Open-ended Scene Graph Generation task [85] in Sec. 5.3. The results and analyses of our proposed CRPE are presented in Sec. 5.4. Note that we utilize a consistent checkpoint for all evaluations.

5.1   Results on General Benchmarks

To evaluate the general ability of ASMv2, we perform a comprehensive comparison with leading MLLMs in Tab. 1. Benefiting from the stronger relation comprehension ability, ASMv2 exhibits SoTA performance on these benchmarks.

Results of Visual Question Answering. On general VQA benchmarks, such as VQAv2 [21] and GQA [25], our ASMv2 demonstrates superior overall performance compared to LLaVA-1.5 [50] and VILA [45]. On the VQAv2 dataset, our ASMv2 outperforms the LLaVA-1.5-13B by 1.0 points. Besides, our model also achieves competitive performance with baselines on text-oriented VQA benchmarks, including VizWiz-VQA [23] and TextVQA [73].

Results of Multi-modal benchmarks. In recent comprehensive benchmarks, which consist of a wide range of sub-tasks covering various fine-grained capabilities, our model significantly outperforms the current SoTA MLLMs, such as

Table 1: Results on 12 general visual-language benchmarks. Benchmark names are abbreviated due to space limits. VQA-v2 [21]; GQA [25]; VizWiz [23]; SQA^I: ScienceQA-IMG [57]; VQA^T: TextVQA [73]; POPE [41]; MME [18]; MMB: MM-Bench [52]; MMB^{CN}: MMBench-Chinese [52]; SEED: SEED-Bench [34]; LLaVA^W: LLaVA-Bench (In-the-Wild) [51]; MM-Vet [91]. *The training images of the datasets are observed during training. The best performances are marked **bold**.

Model	VQA^{v2}	GQA	VizWiz	SQA^I	VQA^T	POPE	MME	MMB	MMB^{CN}	SEED	LLaVA^W	MM-Vet
BLIP-2 [36]	41.0	41.0	19.6	61.0	42.5	85.3	1293.8	-	-	46.4	38.1	22.4
InstructBLIP-7B [15]	-	49.2	34.5	60.5	50.1	-	-	36.0	23.7	53.4	60.9	26.2
InstructBLIP-13B [15]	-	49.5	33.4	63.1	50.7	78.9	1212.8	-	-	-	58.2	25.6
Shikra [8]	77.4*	-	-	-	-	-	-	58.8	-	-	-	-
IDEFICS-9B [28]	50.9	38.4	35.5	-	25.9	-	-	48.2	25.2	-	-	-
IDEFICS-80B [28]	60.0	45.2	36.0	-	30.9	-	-	54.5	38.1	-	-	-
Qwen-VL [2]	78.8*	59.3*	35.2	67.1	63.8	-	-	38.2	7.4	56.3	-	-
Qwen-VL-Chat [2]	78.2*	57.5*	38.9	68.2	61.5	-	1487.5	60.6	56.7	58.2	-	-
LLaVA-1.5-7B [50]	78.5*	62.0*	50.0	66.8	58.2	85.9	1510.7	64.3	58.3	58.6	63.4	30.5
LLaVA-1.5-13B [50]	80.0*	63.3*	53.6	71.6	61.3	85.9	1531.3	67.7	63.6	61.6	70.7	35.4
VILA-7B [45]	79.9*	62.3*	57.8	68.2	64.4	85.5	1533.0	68.9	61.7	61.1	69.7	34.9
VILA-13B [45]	80.8*	63.3*	**60.6**	73.7	**66.6**	84.2	1570.1	70.3	64.3	62.8	73.0	38.8
ASMv2 (ours)	**81.0***	**63.9***	58.1	**87.1***	60.2	**86.3**	**1621.0**	**74.4**	64.3	66.3	**78.9**	**41.3**

Table 2: Accuracy scores on the Referring Expression Comprehension task.

Model	RefCOCO			RefCOCO+			RefCOCOg		Avg.
	Val	Test-A	Test-B	Val	Test-A	Test-B	Val	Test	
OFA-L [80]	79.96	83.67	76.39	68.29	76.00	61.75	67.57	67.50	72.64
VisionLLM-H [82]	-	86.70	-	-	-	-	-	-	-
Shikra-7B [8]	87.01	90.61	80.24	81.60	87.36	72.12	82.27	82.19	82.93
Shikra-13B [8]	87.83	91.11	81.81	82.89	87.79	74.41	84.64	83.16	84.21
Qwen-VL-7B [2]	88.55	92.27	84.51	82.82	88.59	76.79	85.96	86.32	85.73
MiniGPT-V2-7B [7]	88.06	91.29	84.30	79.58	85.52	73.32	84.19	84.31	83.82
Ferret-7B [86]	87.49	91.35	82.45	80.78	87.38	73.14	83.93	84.76	83.91
Ferret-13B [86]	89.48	92.41	84.36	82.81	88.14	75.17	85.83	86.34	85.57
ASMv2 (ours)	**90.56**	**94.24**	**86.24**	**84.81**	**90.83**	**76.89**	**87.52**	**88.26**	**87.42**

LLaVA-1.5 [50] and VILA [45]. Specifically, our model achieves an overall score of 74.4 on MMBench and 1621.0 on MME, surpassing VILA by 4.1 points and 50.9 points separately. Besides, ASMv2 also exhibits state-of-the-art performance on SEED [34], LLaVA-Bench [51], and MM-Vet [91], outperforming baselines by a large margin. These results demonstrate the general ability of our model.

5.2 Results on Region-level Benchmarks

To evaluate the region comprehension and grounding capability, we evaluate ASMv2 on three representative region-level tasks, including (1) Referring Expression Comprehension [31, 58], which requires the model to localize the target object conditioned on the given description. (2) Region Captioning [32, 58], which requires the model to generate a caption for a certain object in the image conditioned on the given region. (3) Referring Question Answering [92], which contains region referring in both questions and answers.

Table 3: Results on the Region Captioning task. We mark the best performance **bold** and the second-best <u>underlined</u>.

Model	VG [32]		RefCOCOg [58]	
	METEOR	CIDEr	METEOR	CIDEr
GRiT [83]	17.1	142.0	15.2	71.6
SLR [89]	-	-	15.4	59.2
SLR+Rerank [89]	-	-	15.9	66.2
Kosmos-2 [63]	-	-	14.1	62.3
GPT4RoI-7B [99]	17.4	145.2	-	-
GPT4RoI-13B [99]	17.6	146.8	-	-
ASM-FT [81]	**18.3**	<u>148.7</u>	**21.8**	<u>107.8</u>
ASMv2 (ours)	<u>17.9</u>	**153.5**	<u>21.7</u>	**114.7**

Table 4: Results on Visual Commonsense Reasoning. Q, A, and R denote the **Q**uestion, **A**nswer, and **R**ationale. X→Y means that the model needs to select the correct option for Y conditioned on X. *The single-task fine-tuning setting.

Method	Validation Acc. (%)		
	Q→A	QA→R	Q→AR
ViLBERT [56]	72.4	74.5	54.0
Unicoder-VL [35]	72.6	74.5	54.5
VLBERT [74]	75.5	77.9	58.9
ERNIE-ViL-L [87]	78.5	83.4	65.8
VILLA [19]	78.5	82.6	65.2
*GPT4RoI-7B [99]	87.4	89.6	78.6
ASMv2 (ours)	87.8	88.8	78.4
*ASMv2 (ours)	**88.4**	**89.9**	**79.4**

Results of Referring Expression Comprehension. Our ASMv2 achieves state-of-the-art performance on the representative REC benchmarks [31,58]. As shown in Tab. 2, our ASMv2 significantly outperforms current state-of-the-art MLLMs, including Qwen-VL [2] and Ferret [86]. Specifically, our model surpasses Ferret by an average of 1.85 points across all benchmarks.

Results of Region Captioning. Our model demonstrates state-of-the-art performance on the representative region captioning benchmarks, including VG [32] and RefCOCOg [58]. As shown in Tab. 3, our model achieves a CIDEr score of 114.7 on RefCOCOg, which surpasses the current state-of-the-art model (*i.e.* ASM-FT) by 6.9 points. On the VG dataset, our model also exhibits competitive results compared to the current state-of-the-art model.

Results of Referring Question Answering. We evaluate the Referring Question Answering (RQA) ability of ASMv2 on the Visual Commonsense Reasoning (VCR) dataset [92], which evaluates the commonsense reasoning abilities in the form of single-choice questions. The questions and candidate choices in VCR contain region referring. The results are presented in Tab. 4. Although trained

in a multi-task setting, ASMv2 exhibits competitive performance compared to the current state-of-the-art model (*i.e.*, GPT4RoI [99]), which is finetuned on the VCR dataset in a single task setting. In addition, after the single task finetuning, our model outperforms GPT4RoI by 0.8 points.

5.3   Results on Open-ended Scene Graph Generation

In this section, we evaluate the Relation Conversation capability of our model through the Open-ended Scene Graph Generation task on the Panoptic Scene Graph (PSG) dataset [85], which is a widely-used benchmark for the scene graph generation. See Appendix C.2 for the results on the Predicate Classification task.

Baselines. Despite the powerful performance, most previous methods [79, 85, 105, 106] are constrained by pre-defined label sets and struggle to capture a diverse range of visual concepts from natural language in an open-ended manner. On the other hand, TextPSG [102] explores a methodology for generating scene graphs in an open-ended manner, which first generates the region proposals and then asks BLIP [37] to predict the semantic tags and predicate labels for these regions auto-regressively. Here, we consider traditional close-set scene graph generation models and TextPSG as our baseline in OpenSGG.

Metrics. Following the common practice [85, 102], we report the triplet Recall and mean Recall for every predicate category (mRecall) in the OpenSGG task. Concretely, a scene graph consists of a set of triplets (`subject, predicate, object`). A triplet is considered to be correct if the phrase labels are all correct and the location of the subject and object should match the ground truth with IoU greater than 0.5 respectively. We also report #Tuples to denote the average number of predicted tuples for each generated scene graph.

Results. As shown in Tab. 5, our ASMv2 demonstrates state-of-the-art performance in the OpenSGG task. Specifically, our ASMv2 significantly outperforms TextPSG by 8.7 points in recall while generating a significantly fewer average number of tuples compared to it (9.2 *vs.* 100.0). Note that having more tuples generally implies an advantage in computing recall. When compared to traditional scene graph generation models, which generate scene graphs in a close-ended manner, ASMv2 also exhibits competitive performance. Despite generating fewer tuples, our model maintains a competitive recall of 14.2 and a mean recall of

Table 5: Recall scores on PSG. Gray denotes that the model generates the scene graphs in a close-ended manner.

Model	#Tuples	Recall	mRecall
IMP [84]	20.0	16.5	6.5
MOTIFS [94]	20.0	20.0	9.1
VCTree [76]	20.0	20.6	9.7
GPSNet [47]	20.0	17.8	7.0
PSGFormer [85]	20.0	18.6	16.7
TextPSG [102]	50.0	4.8	-
TextPSG [85]	100.0	5.5	-
ASMv2 (ours)	9.2	14.2	10.3

10.3. Another factor negatively impacting the performance is that our ASMv2 generates scene graphs in an open-ended manner while recall is calculated in an exact-match manner. Therefore, the triplets (people, standing on, grass) and (person, standing on, grass) are considered mismatched even though they represent the same semantics. A more appropriate metric for the OpenSGG task will be left for future work.

5.4   Results on CRPE

In this section, we evaluate the relation comprehension ability of our ASMv2 and current leading MLLMs [2,50] using our proposed CRPE benchmark. This bench-

Table 6: Accuracy scores on CRPE.

Model	Exist.	Subj.	Pred.	Obj.	Overall
Qwen-VL [2]	85.11	45.66	38.19	31.60	38.48
LLaVA-1.5 [50]	88.69	57.44	54.24	55.21	55.63
ASMv2 (ours)	**92.14**	**69.21**	**58.95**	**65.34**	**64.50**

mark consists of four splits: Existence, Subject, Predicate, and Object. The Existence split evaluates the models' object recognition ability while the remaining splits are designed to evaluate the models' relation comprehension ability. In addition to reporting the performance of each split in the benchmark individually, we also report the average score of the latter three splits as the overall score for relation comprehension ability.

As shown in Tab. 6, the performance of existing MLLMs on the Existence questions is significantly higher than on the Subject, Predicate, and Object questions. This suggests that these models have a more robust capability to recognize objects within an image than to comprehend the relations between them. Specifically, our ASMv2 shows a remarkable improvement in understanding object relations compared to the other models. For example, ASMv2 achieves an overall accuracy of 64.50, which is significantly higher than the 55.63 of LLaVA-1.5 and the 38.48 of Qwen-VL. These results demonstrate that our model can comprehend the relations between the objects within the image better, benefiting from the training of relation conversation data.

6   Conclusion

In this paper, we propose a novel task, termed Relation Conversation (ReC), to challenge the model to understand the relations between the objects within the image. We construct the All-Seeing Dataset V2 (AS-V2), which is a high-quality ReC dataset to unlock the ReC ability of Multi-modal Large Language Models (MLLMs) and the CRPE to quantitatively evaluate the relation comprehension ability. Leveraging AS-V2 and other general multimodal corpora for training, we introduce the All-Seeing Model v2 (ASMv2), which exhibits stronger relation comprehension ability compared to existing leading MLLMs and achieves state-of-the-art performance on the Open-ended Scene Graph Generation task and various general image-level and region-level tasks. We hope that our work can inspire more future research and contribute to the evolution towards artificial general intelligence, equipping artificial intelligence systems with an "all-seeing eye" to achieve a deeper understanding of the world.

References

1. Andreas, J., Rohrbach, M., Darrell, T., Klein, D.: Neural module networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 39–48 (2016) 5

2. Bai, J., Bai, S., Yang, S., Wang, S., Tan, S., Wang, P., Lin, J., Zhou, C., Zhou, J.: Qwen-vl: A frontier large vision-language model with versatile abilities. arXiv preprint arXiv:2308.12966 (2023) 2, 4, 11, 12, 14

3. Betker, J., Goh, G., Jing, L., Brooks, T., Wang, J., Li, L., Ouyang, L., Zhuang, J., Lee, J., Guo, Y., et al.: Improving image generation with better captions. Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf (2023) 8

4. 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) 10, 27

5. 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) 4, 28

6. Caesar, H., Uijlings, J., Ferrari, V.: Coco-stuff: Thing and stuff classes in context. In: CVPR (2018) 5

7. Chen, J., Zhu, D., Shen, X., Li, X., Liu, Z., Zhang, P., Krishnamoorthi, R., Chandra, V., Xiong, Y., Elhoseiny, M.: Minigpt-v2: large language model as a unified interface for vision-language multi-task learning. arXiv preprint arXiv:2310.09478 (2023) 11

8. Chen, K., Zhang, Z., Zeng, W., Zhang, R., Zhu, F., Zhao, R.: Shikra: Unleashing multimodal llm's referential dialogue magic. arXiv preprint arXiv:2306.15195 (2023) 2, 4, 11

9. 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) 10, 27

10. Chen, X., Fang, H., Lin, T.Y., Vedantam, R., Gupta, S., Dollár, P., Zitnick, C.L.: Microsoft coco captions: Data collection and evaluation server. arXiv preprint arXiv:1504.00325 (2015) 3, 5

11. Chen, Z., Wu, J., Wang, W., Su, W., Chen, G., Xing, S., Zhong, M., Zhang, Q., Zhu, X., Lu, L., Li, B., Luo, P., Lu, T., Qiao, Y., Dai, J.: Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. arXiv preprint arXiv:2312.14238 (2023) 1, 4

12. Clark, C., Gardner, M.: Simple and effective multi-paragraph reading comprehension. arXiv preprint arXiv:1710.10723 (2017) 10, 27

13. Cong, Y., Yang, M.Y., Rosenhahn, B.: Reltr: Relation transformer for scene graph generation. IEEE Transactions on Pattern Analysis and Machine Intelligence (2023) 5

14. Dai, B., Zhang, Y., Lin, D.: Detecting visual relationships with deep relational networks. In: CVPR (2017) 5

15. Dai, W., Li, J., Li, D., Huat, A., Zhao, J., Wang, W., Li, B., Fung, P., Hoi, S.: Instructblip: Towards general-purpose vision-language models with instruction tuning. arXiv preprint arXiv:2305.06500 (2023) 11

16. Dupty, M.H., Zhang, Z., Lee, W.S.: Visual relationship detection with low rank non-negative tensor decomposition. In: AAAI (2020) 5

17. Fang, Y., Wang, W., Xie, B., Sun, Q., Wu, L., Wang, X., Huang, T., Wang, X., Cao, Y.: Eva: Exploring the limits of masked visual representation learning at scale. arXiv preprint arXiv:2211.07636 (2022) 4

18. 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) 4, 11, 29

19. Gan, Z., Chen, Y.C., Li, L., Zhu, C., Cheng, Y., Liu, J.: Large-scale adversarial training for vision-and-language representation learning. NeurIPS (2020) 12

20. Gao, P., Geng, S., Zhang, R., Ma, T., Fang, R., Zhang, Y., Li, H., Qiao, Y.: Clip-adapter: Better vision-language models with feature adapters. arXiv preprint arXiv:2110.04544 (2021) 8

21. 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) 10, 11, 27

22. Gu, J., Zhao, H., Lin, Z., Li, S., Cai, J., Ling, M.: Scene graph generation with external knowledge and image reconstruction. In: CVPR (2019) 5

23. Gurari, D., Li, Q., Stangl, A.J., Guo, A., Lin, C., Grauman, K., Luo, J., Bigham, J.P.: Vizwiz grand challenge: Answering visual questions from blind people. In: CVPR (2018) 10, 11

24. Hu, Y., Chen, S., Chen, X., Zhang, Y., Gu, X.: Neural message passing for visual relationship detection. In: ICMLW (2019) 5

25. Hudson, D.A., Manning, C.D.: Gqa: A new dataset for real-world visual reasoning and compositional question answering. In: CVPR (2019) 10, 11, 27

26. Hung, Z.S., Mallya, A., Lazebnik, S.: Contextual translation embedding for visual relationship detection and scene graph generation. IEEE transactions on pattern analysis and machine intelligence pp. 3820–3832 (2020) 5

27. Hwang, S.J., Ravi, S.N., Tao, Z., Kim, H.J., Collins, M.D., Singh, V.: Tensorize, factorize and regularize: Robust visual relationship learning. In: CVPR (2018) 5

28. IDEFICS: Introducing idefics: An open reproduction of state-of-the-art visual language model. https://huggingface.co/blog/idefics (2023) 11

29. Jia, C., Yang, Y., Xia, Y., Chen, Y.T., Parekh, Z., Pham, H., Le, Q., Sung, Y.H., Li, Z., Duerig, T.: Scaling up visual and vision-language representation learning with noisy text supervision. In: Int. Conf. Mach. Learn. (2021) 4

30. 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) 5, 10, 27

31. Kazemzadeh, S., Ordonez, V., Matten, M., Berg, T.: Referitgame: Referring to objects in photographs of natural scenes. In: EMNLP (2014) 2, 4, 10, 11, 12, 27, 29

32. 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) 2, 5, 10, 11, 12, 27

33. Lai, X., Tian, Z., Chen, Y., Li, Y., Yuan, Y., Liu, S., Jia, J.: Lisa: Reasoning segmentation via large language model. arXiv preprint arXiv:2308.00692 (2023) 4

34. Li, B., Wang, R., Wang, G., Ge, Y., Ge, Y., Shan, Y.: Seed-bench: Benchmarking multimodal llms with generative comprehension. arXiv preprint arXiv:2307.16125 (2023) 11

35. Li, G., Duan, N., Fang, Y., Gong, M., Jiang, D.: Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. In: AAAI (2020) 12

36. Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pretraining with frozen image encoders and large language models. arXiv preprint arXiv:2301.12597 (2023) 4, 11

37. Li, J., Li, D., Xiong, C., Hoi, S.: Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In: ICML (2022) 10, 13

38. Li, J., Selvaraju, R., Gotmare, A., Joty, S., Xiong, C., Hoi, S.C.H.: Align before fuse: Vision and language representation learning with momentum distillation. NeurIPS (2021) 4

39. Li, K., He, Y., Wang, Y., Li, Y., Wang, W., Luo, P., Wang, Y., Wang, L., Qiao, Y.: Videochat: Chat-centric video understanding. arXiv preprint arXiv:2305.06355 (2023) 4

40. Li, R., Zhang, S., He, X.: Sgtr: End-to-end scene graph generation with transformer. In: CVPR (2022) 5

41. Li, Y., Du, Y., Zhou, K., Wang, J., Zhao, W.X., Wen, J.R.: Evaluating object hallucination in large vision-language models. EMNLP (2023) 11

42. Li, Y., Ouyang, W., Wang, X., Tang, X.: Vip-cnn: Visual phrase guided convolutional neural network. In: CVPR (2017) 5

43. Li, Y., Ouyang, W., Zhou, B., Wang, K., Wang, X.: Scene graph generation from objects, phrases and region captions. In: ICCV (2017) 5

44. Liao, W., Rosenhahn, B., Shuai, L., Ying Yang, M.: Natural language guided visual relationship detection. In: CVPRW (2019) 5

45. Lin, J., Yin, H., Ping, W., Lu, Y., Molchanov, P., Tao, A., Mao, H., Kautz, J., Shoeybi, M., Han, S.: Vila: On pre-training for visual language models. arXiv preprint arXiv:2312.07533 (2023) 10, 11

46. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: ECCV (2014) 3, 5, 7

47. Lin, X., Ding, C., Zeng, J., Tao, D.: Gps-net: Graph property sensing network for scene graph generation. In: CVPR. pp. 3746–3753 (2020) 13, 28

48. Liu, F., Emerson, G., Collier, N.: Visual spatial reasoning. Transactions of the Association for Computational Linguistics (2023) 5, 10

49. Liu, F., Lin, K., Li, L., Wang, J., Yacoob, Y., Wang, L.: Aligning large multimodal model with robust instruction tuning. arXiv preprint arXiv:2306.14565 (2023) 10, 27

50. Liu, H., Li, C., Li, Y., Lee, Y.J.: Improved baselines with visual instruction tuning. arXiv preprint arXiv:2310.03744 (2023) 1, 4, 8, 10, 11, 14

51. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. arXiv preprint arXiv:2304.08485 (2023) 1, 4, 10, 11, 27

52. Liu, Y., Duan, H., Zhang, Y., Li, B., Zhang, S., Zhao, W., Yuan, Y., Wang, J., He, C., Liu, Z., et al.: Mmbench: Is your multi-modal model an all-around player? arXiv preprint arXiv:2307.06281 (2023) 4, 7, 11

53. Liu, Z., He, Y., Wang, W., Wang, W., Wang, Y., Chen, S., Zhang, Q., Lai, Z., Yang, Y., Li, Q., Yu, J., et al.: Interngpt: Solving vision-centric tasks by interacting with chatgpt beyond language. arXiv preprint arXiv:2305.05662 (2023) 4

54. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: ICLR (2019) 27

55. Lu, C., Krishna, R., Bernstein, M., Fei-Fei, L.: Visual relationship detection with language priors. In: ECCV (2016) 4, 5

56. Lu, J., Batra, D., Parikh, D., Lee, S.: Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. NeurIPS (2019) 12

57. 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) 10, 11, 27

58. Mao, J., Huang, J., Toshev, A., Camburu, O., Yuille, A.L., Murphy, K.: Generation and comprehension of unambiguous object descriptions. In: CVPR (2016) 2, 4, 10, 11, 12, 27, 29

59. Marino, K., Rastegari, M., Farhadi, A., Mottaghi, R.: Ok-vqa: A visual question answering benchmark requiring external knowledge. In: CVPR (2019) 27

60. 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) 27

61. OpenAI: Gpt-4 technical report. arXiv preprint arXiv:2303.08774 (2023) 4

62. OpenAI: Gpt-4v(ision) system card (2023) 4, 5

63. Peng, Z., Wang, W., Dong, L., Hao, Y., Huang, S., Ma, S., Wei, F.: Kosmos-2: Grounding multimodal large language models to the world. arXiv preprint arXiv:2306.14824 (2023) 2, 4, 10, 12

64. Plummer, B.A., Wang, L., Cervantes, C.M., Caicedo, J.C., Hockenmaier, J., Lazebnik, S.: Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In: ICCV (2015) 10

65. Qi, M., Li, W., Yang, Z., Wang, Y., Luo, J.: Attentive relational networks for mapping images to scene graphs. In: CVPR (2019) 5

66. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: ICML (2021) 4

67. Rasheed, H., Maaz, M., Shaji, S., Shaker, A., Khan, S., Cholakkal, H., Anwer, R.M., Xing, E., Yang, M.H., Khan, F.S.: Glamm: Pixel grounding large multimodal model. arXiv preprint arXiv:2311.03356 (2023) 4

68. Schwenk, D., Khandelwal, A., Clark, C., Marino, K., Mottaghi, R.: A-okvqa: A benchmark for visual question answering using world knowledge. In: ECCV (2022) 27

69. ShareGPT: https://sharegpt.com/ (2023) 27

70. Sharma, P., Ding, N., Goodman, S., Soricut, R.: Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In: ACL (2018) 10

71. Shit, S., Koner, R., Wittmann, B., Paetzold, J., Ezhov, I., Li, H., Pan, J., Sharifzadeh, S., Kaissis, G., Tresp, V., et al.: Relationformer: A unified framework for image-to-graph generation. In: ECCV (2022) 5

72. Sidorov, O., Hu, R., Rohrbach, M., Singh, A.: Textcaps: a dataset for image captioning with reading comprehension. In: ECCV. pp. 742–758. Springer (2020) 27

73. Singh, A., Natarajan, V., Shah, M., Jiang, Y., Chen, X., Batra, D., Parikh, D., Rohrbach, M.: Towards vqa models that can read. In: CVPR (2019) 10, 11

74. Su, W., Zhu, X., Cao, Y., Li, B., Lu, L., Wei, F., Dai, J.: Vl-bert: Pre-training of generic visual-linguistic representations. arXiv preprint arXiv:1908.08530 (2019) 12

75. Suhr, A., Lewis, M., Yeh, J., Artzi, Y.: A corpus of natural language for visual reasoning. In: Ann. Meeting of the Assoc. for Comput. Linguistics (2017) 5

76. 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) 13, 28

77. Tian, C., Zhu, X., Xiong, Y., Wang, W., Chen, Z., Wang, W., Chen, Y., Lu, L., Lu, T., Zhou, J., et al.: Mm-interleaved: Interleaved image-text generative modeling via multi-modal feature synchronizer. arXiv preprint arXiv:2401.10208 (2024) 4

78. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al.: Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971 (2023) 4

79. Wang, J., Wen, Z., Li, X., Guo, Z., Yang, J., Liu, Z.: Pair then relation: Pair-net for panoptic scene graph generation. arXiv preprint arXiv:2307.08699 (2023) 13

80. Wang, P., Yang, A., Men, R., Lin, J., Bai, S., Li, Z., Ma, J., Zhou, C., Zhou, J., Yang, H.: Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In: ICML (2022) 11

81. 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) 2, 4, 10, 12, 27

82. Wang, W., Chen, Z., Chen, X., Wu, J., Zhu, X., Zeng, G., Luo, P., Lu, T., Zhou, J., Qiao, Y., et al.: Visionllm: Large language model is also an open-ended decoder for vision-centric tasks. arXiv preprint arXiv:2305.11175 (2023) 2, 4, 11

83. Wu, J., Wang, J., Yang, Z., Gan, Z., Liu, Z., Yuan, J., Wang, L.: Grit: A generative region-to-text transformer for object understanding. arXiv preprint arXiv:2212.00280 (2022) 12

84. Xu, D., Zhu, Y., Choy, C.B., Fei-Fei, L.: Scene graph generation by iterative message passing. In: CVPR (2017) 13, 28

85. Yang, J., Ang, Y.Z., Guo, Z., Zhou, K., Zhang, W., Liu, Z.: Panoptic scene graph generation. In: ECCV (2022) 3, 5, 7, 10, 13, 27, 28

86. You, H., Zhang, H., Gan, Z., Du, X., Zhang, B., Wang, Z., Cao, L., Chang, S.F., Yang, Y.: Ferret: Refer and ground anything anywhere at any granularity. arXiv preprint arXiv:2310.07704 (2023) 11, 12

87. Yu, F., Tang, J., Yin, W., Sun, Y., Tian, H., Wu, H., Wang, H.: Ernie-vil: Knowledge enhanced vision-language representations through scene graphs. In: AAAI (2021) 12

88. Yu, J., Wang, Z., Vasudevan, V., Yeung, L., Seyedhosseini, M., Wu, Y.: Coca: Contrastive captioners are image-text foundation models. arXiv preprint arXiv:2205.01917 (2022) 4

89. Yu, L., Tan, H., Bansal, M., Berg, T.L.: A joint speaker-listener-reinforcer model for referring expressions. In: CVPR (2017) 12

90. Yu, Q., Sun, Q., Zhang, X., Cui, Y., Zhang, F., Wang, X., Liu, J.: Capsfusion: Rethinking image-text data at scale. arXiv preprint arXiv:2310.20550 (2023) 4

91. 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) 11, 29

92. Zellers, R., Bisk, Y., Farhadi, A., Choi, Y.: From recognition to cognition: Visual commonsense reasoning. In: CVPR (2019) 2, 10, 11, 12, 27, 29

93. Zellers, R., Yatskar, M., Thomson, S., Choi, Y.: Neural motifs: Scene graph parsing with global context. In: CVPR (2018) 5

94. Zellers, R., Yatskar, M., Thomson, S., Choi, Y.: Neural motifs: Scene graph parsing with global context. In: CVPR (2018) 13, 28

95. Zhai, X., Wang, X., Mustafa, B., Steiner, A., Keysers, D., Kolesnikov, A., Beyer, L.: Lit: Zero-shot transfer with locked-image text tuning. In: Proceedings of the

IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18123–18133 (2022) 4

96. Zhang, A., Zhao, L., Xie, C.W., Zheng, Y., Ji, W., Chua, T.S.: Next-chat: An lmm for chat, detection and segmentation. arXiv preprint arXiv:2311.04498 (2023) 4

97. Zhang, H., Kyaw, Z., Chang, S.F., Chua, T.S.: Visual translation embedding network for visual relation detection. In: CVPR (2017) 5

98. Zhang, J., Kalantidis, Y., Rohrbach, M., Paluri, M., Elgammal, A., Elhoseiny, M.: Large-scale visual relationship understanding. In: AAAI (2019) 5

99. Zhang, S., Sun, P., Chen, S., Xiao, M., Shao, W., Zhang, W., Chen, K., Luo, P.: Gpt4roi: Instruction tuning large language model on region-of-interest. arXiv preprint arXiv:2307.03601 (2023) 4, 12, 13

100. 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) 10, 27

101. Zhao, B., Wu, B., Huang, T.: Svit: Scaling up visual instruction tuning. arXiv preprint arXiv:2307.04087 (2023) 10, 27

102. Zhao, C., Shen, Y., Chen, Z., Ding, M., Gan, C.: Textpsg: Panoptic scene graph generation from textual descriptions. In: ICCV (2023) 13

103. Zheng, L., Chiang, W.L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E.P., Zhang, H., Gonzalez, J.E., Stoica, I.: Judging llm-as-a-judge with mt-bench and chatbot arena (2023) 8

104. Zheng, S., Chen, S., Jin, Q.: Visual relation detection with multi-level attention. In: ACM MM (2019) 5

105. Zhong, Y., Shi, J., Yang, J., Xu, C., Li, Y.: Learning to generate scene graph from natural language supervision. In: ICCV (2021) 13

106. Zhou, Z., Shi, M., Caesar, H.: Vlprompt: Vision-language prompting for panoptic scene graph generation. arXiv preprint arXiv:2311.16492 (2023) 5, 13

107. Zhu, D., Chen, J., Shen, X., Li, X., Elhoseiny, M.: Minigpt-4: Enhancing vision-language understanding with advanced large language models. arXiv preprint arXiv:2304.10592 (2023) 4

108. 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) 27
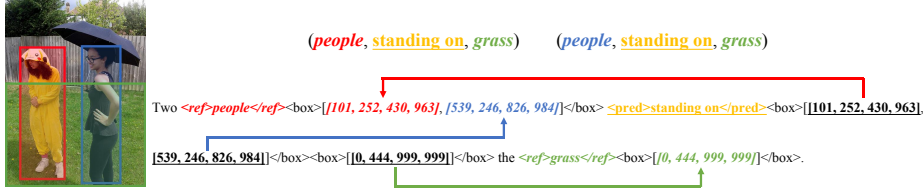
Fig. 6: A complex example for the formulation of Relation Conversation.

A Relation Conversation

In this section, we introduce more details about the formulation of Relation Conversation. As depicted in Fig. 6, when a certain text span is associated with multiple regions, these bounding boxes are formatted as:

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

where $[x_1^i, y_1^i, x_2^i, y_2^i]$ 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. 6, 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. 7 to 9. Besides, prompts used to generate detailed description data are shown in Tabs. 7 and 8.
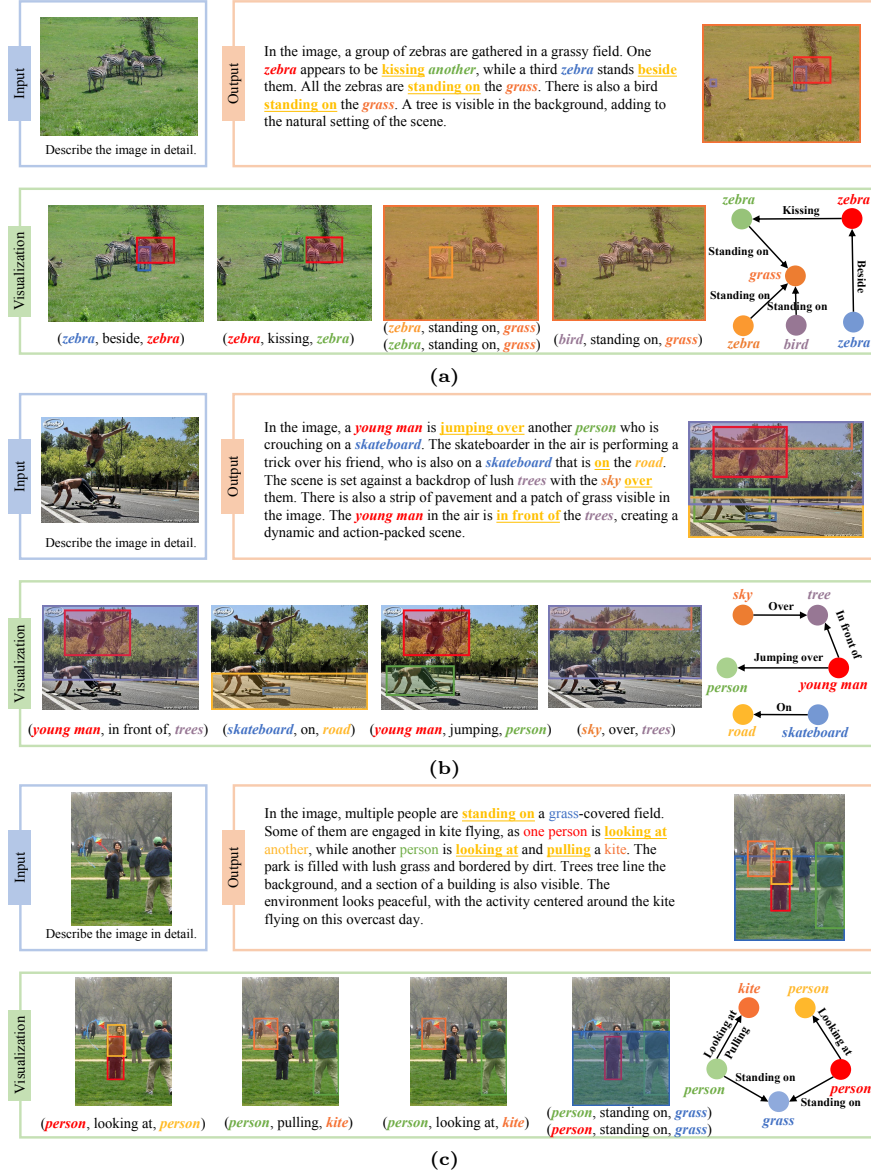
Fig. 7: Data Examples of Detailed Description task in AS-V2.

Fig. 8: Data Examples of Region Captioning task in AS-V2.

Fig. 9: Data Examples of Conversation task in AS-V2. Due the space limitations, we exhibit only one turn for each conversation.

Table 7: 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 men-
tion 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 8: 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 9: 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 [72], ShareGPT4V [9]
VQA	314K	VQAv2 [21], GQA [25], OKVQA [59], A-OKVQA [68], ScienceQA [57], CLEVR [30], Visual7W [108]
OCR	157K	ST-VQA [4], LLaVAR [100], OCR-VQA [60], DocVQA [12]
Grounding	643K	RefCOCO/+/g [31, 58]
RegionVQA	2.3M	RefCOCOg [58], VG [32], VCR [92], AS-Core [81]
Conversation	500K	LLaVA-Instruct [51], SVIT [101], LRV [49], AS-V2 (ours)
Text	40K	ShareGPT [69]

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 [54] 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 336×336.

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 [85]. 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 10: Recall scores on Predicate Classification task.

Method	Predicate Classification					
	R@20	mR@20	R@50	mR@50	R@100	mR@100
IMP [84]	30.5	9.0	35.9	10.5	38.3	11.3
MOTIFS [94]	45.1	19.9	50.5	21.5	52.5	22.2
VCTree [76]	**45.9**	**21.4**	**51.2**	23.1	**53.1**	23.8
GPSNet [47]	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)
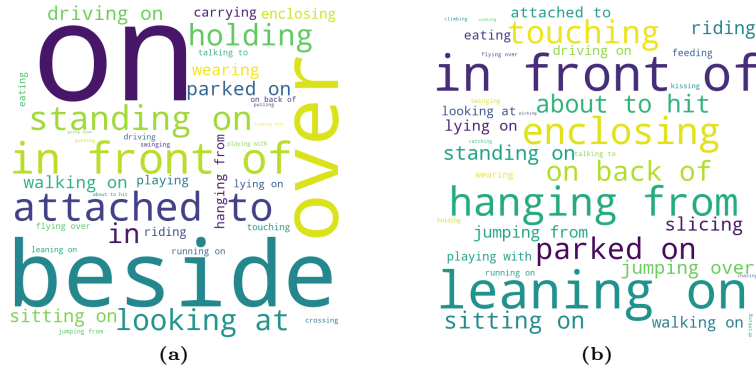
Fig. 10: Word Clouds for evaluation data in PSG. Fig. 10a visualizes the distribution of ground-truth predicates while Fig. 10b visualizes those predicted by ASMv2.

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 [5] to generate the scores for each predicate label. Following prior works [76,85], we report the Recall and mean Recall (mRecall) here.

Results. As shown in Tab. 10, 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. 10, 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 11: Ablation results. We report the $Q \rightarrow AR$ accuracy for VCR [92] and the overall accuracy for CRPE. REC denotes the average accuracy score across Ref-COCO [31], RefCOCO+ [58], and RefCOCOg [58].

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. 11, 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 [18], MM-Vet [91], and Referring Expression Comprehension (REC) Benchmarks [31, 58], 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 [31, 58], VCR [92], and CRPE. The performance on MME [18] 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

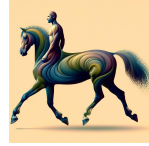In this section, we present more examples of abnormal data in CRPE in Fig. 11.

Fig. 11: Data examples of abnormal data in the CRPE.