LLaVA-SpaceSGG: Visual Instruct Tuning for Open-vocabulary Scene Graph Generation with Enhanced Spatial Relations

Mingjie Xu[1*], Mengyang Wu[2*], Yuzhi Zhao[3†], Jason Chun Lok Li[4], Weifeng Ou[5]
parasolohalo@gmail.com, yzzhao2-c@my.cityu.edu.hk
[1]Independent Researcher [2]The Chinese University of Hong Kong [3]City University of Hong Kong
[4]The University of Hong Kong [5]Dongguan University of Technology

Abstract

Scene Graph Generation (SGG) converts visual scenes into structured graph representations, providing deeper scene understanding for complex vision tasks. However, existing SGG models often overlook essential spatial relationships and struggle with generalization in open-vocabulary contexts. To address these limitations, we propose LLaVA-SpaceSGG, a multimodal large language model (MLLM) designed for open-vocabulary SGG with enhanced spatial relation modeling. To train it, we collect the SGG instruction-tuning dataset, named SpaceSGG. This dataset is constructed by combining publicly available datasets and synthesizing data using open-source models within our data construction pipeline. It combines object locations, object relations, and depth information, resulting in three data formats: spatial SGG description, question-answering, and conversation. To enhance the transfer of MLLMs' inherent capabilities to the SGG task, we introduce a two-stage training paradigm. Experiments show that LLaVA-SpaceSGG outperforms other open-vocabulary SGG methods, boosting recall by 8.6% and mean recall by 28.4% compared to the baseline. Our codebase, dataset, and trained models are publicly accessible on GitHub at the following URL: https://github.com/Endlinc/LLaVA-SpaceSGG.

1. Introduction

Scene Graph Generation (SGG) is a fundamental scene understanding task that involves detecting the entities and predicting their relationships in an image to form a scene graph (see Figure 1 (a) and (b)). The scene graph can be formulated as several text tuples of (subject, predicate, object), where the nodes denote objects and the edges denote relationships between different object pairs, respectively. Since the scene graph is a concise semantic representation of an image, it can be an intermediate feature for complex vision tasks. For instance, it has been applied in diverse downstream tasks such as visual question answering [3, 12, 16, 28], image captioning [9, 27, 30, 46, 58], image retrieval [20, 44, 60], etc.

Recent approaches have attempted to generate scene graphs in a supervised manner, yielding remarkable results. Nonetheless, we have identified two challenges that constrain the overall performance:

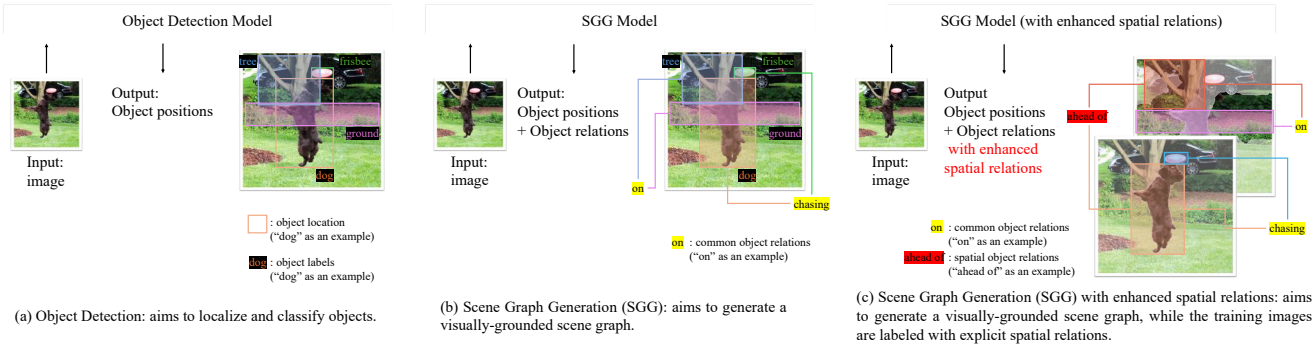
1) Open-vocabulary SGG: Existing SGG methods often require direct supervision with a fixed set of labels and their generalization ability on open-set images is unsatisfactory;

2) Lack of spatial relations: Since existing SGG datasets are mainly annotated on 2D images, the annotation progress mainly focuses on common relationships and neglects 3D spatial relationships between certain objects.

In pursuit of open-vocabulary SGG, recent approaches, exemplified by ASMv2 [51], have integrated state-of-the-art vision-language models like CLIP [42] and LLaVA [32], leveraging rich and diverse training data encompassing various modalities. Nonetheless, these methods tend to overlook the crucial 3D spatial relationships that form integral elements of SGG. To emphasize spatial relations, Pu *et al.* [40] and Li *et al.* [26] integrated spatially specific blocks to assimilate spatial correlations and enhance spatial contextual understanding. However, it does not efficiently balance the original information and new information extracted by proposed blocks.

To address the two challenges simultaneously, we propose LLaVA-SpaceSGG, specifically designed to tackle both open-vocabulary SGG and spatial relationship modeling (see Figure 1 (c)). We have extended the LLaVA-1.5 framework [33, 35] and curated a SpaceSGG instruction-tuning dataset. Then, we introduce a two-stage training paradigm to train the LLaVA-SpaceSGG. In the first stage, we align an image model (e.g., CLIP [42]) with a text model

*Equal Contribution.
†Corresponding Author.



(a) Object Detection: aims to localize and classify objects.

(b) Scene Graph Generation (SGG): aims to generate a visually-grounded scene graph.

(c) Scene Graph Generation (SGG) with enhanced spatial relations: aims to generate a visually-grounded scene graph, while the training images are labeled with explicit spatial relations.

Figure 1. The illustration of different tasks: (a) Object Detection, (b) Scene Graph Generation (SGG), and (c) Scene Graph Generation (SGG) with enhanced spatial relations. By additionally leveraging spatial relationships, we propose the LLaVA-SpaceSGG framework.

(e.g., Llama 2 [50]), enabling the model to excel in open-vocabulary SGG, leveraging the vast pre-training datasets. In the second stage, we refine the model's comprehension of region-level spatial relationships, crucial for SGG. This dual-phase approach comprises a pre-training stage succeeded by an instruction-tuning phase, akin to ASMv2 [51]. To further enhance spatial understanding, we fully exploit SGG-related data in the second instruction-tuning phase, incorporating both the general SGG instruction-tuning dataset from [51] and our newly created SpaceSGG dataset.

Our dataset introduces two key improvements. Firstly, it captures both plane and depth coordinates, enriching the spatial relationships (such as front-back relationships) between objects. Specifically, we first use a depth estimation algorithm [57] to generate a depth map from an image, then construct a 3D scene by [14], and finally extract 3D SGG from the 3D scene. Secondly, our dataset generates three distinct data formats: spatial descriptions (SpaceSGG-Desc), single-turn question answering (SpaceSGG-QA), and multi-turn conversations (SpaceSGG-Conv) to enhance the model's spatial reasoning capabilities. SpaceSGG-Desc includes both plane and depth SGG descriptions. SpaceSGG-QA emphasizes the spatial relationships between two objects by depth comparison and multi-view questions. SpaceSGG-Conv contains the complete reasoning process from an image to SGG based on chain-of-thought (CoT) multi-turn dialogue [53].

To evaluate the ability of the proposed LLaVA-SpaceSGG, we conduct experiments on a general Panoptic Scene Graph dataset (PSG) [55]. To further examine the spatial understanding ability, we construct a spatial relation validation dataset, which contains 271 labeled question-answer pairs on the COCO dataset [30]. Our LLaVA-SpaceSGG outperforms current state-of-the-art models by 8.6% recall and by 28.4% mean recall in the PSG validation set. It also outperforms existing methods with respect to an accuracy of 3.8% in the proposed spatial relation validation set. Experiments show that LLaVA-SpaceSGG is able to discover, map, and predict richer spatial relationships while others do not. And it demonstrates that the SpaceSGG dataset greatly contributes to improving the model's understanding of spatial relationships.

In summary, there are three main contributions:

1) To enhance spatial understanding in SGG, we collect the SpaceSGG dataset, along with a novel data generation pipeline. This dataset integrates both 2D and 3D scene information, resulting in a more comprehensive representation of object relations that captures spatial context, object positions, and depth. This fusion addresses critical limitations in existing SGG datasets, which often lack detailed spatial information.

2) Utilizing the SpaceSGG dataset, we develop LLaVA-SpaceSGG, a multimodal large language model designed specifically for open-vocabulary SGG tasks. In order to enhance the adaptation of MLLMs to the SGG domain, we propose a task-specific two-stage training strategy. This methodology notably enhances the model's capacity to interpret spatial relationships within intricate visual contexts.

3) The LLaVA-SpaceSGG model showcases the state-of-the-art performance on the well-established PSG validation set, surpassing current methods in recall and mean recall. Furthermore, to evaluate the model's proficiency in abstracting spatial relationships, we present a novel spatial relation validation set. Our model attains reliable and consistent performance on this new benchmark, underscoring its efficacy in capturing spatial dynamics.

2. Related Work

2.1. Scene Graph Generation (SGG)

SGG has become increasingly important in the computer vision area. The interest in this field was initially sparked by Lu *et al*. [35], which focuses on the relationships between objects that can be abstractly represented by a graph of nodes and edges. Such scene graphs are immensely helpful for models to understand interactions among objects within

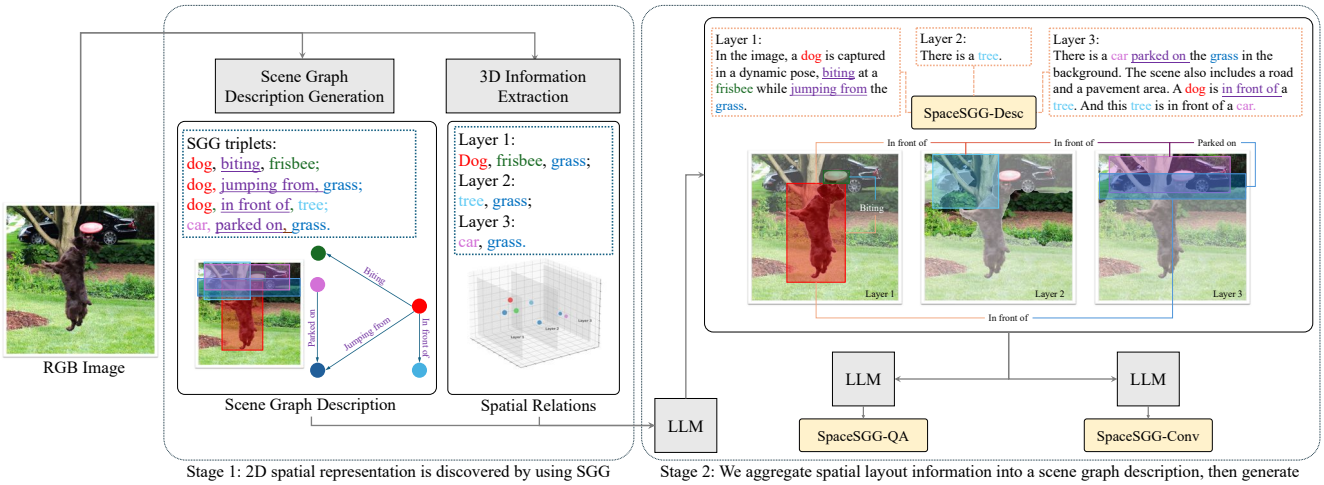


Figure 2. SpaceSGG dataset construction pipeline. We utilize both SGG description and spatial relationships, where we generate 3 types of data: spatial scene detailed descriptions (SpaceSGG-Desc), QA (SpaceSGG-QA), and multi-turn conversations (SpaceSGG-Conv).

images. However, early works [10, 27, 54, 56] in this field overly simplified the scenes used for training, containing only a few objects and thus leading to an over-concentration of object relationships, which hindered the model's ability to learn generalized representational knowledge. Subsequently, Yang *et al*. [55] proposed conducting a comprehensive segmentation of the entire image and attempted to identify all relevant relationships between the resulting segments, thereby enriching the complexity of objects and their relationships. We aim to improve the performance of SGG by focusing on a fundamental yet underexplored aspect: spatial relationships. These relationships, which naturally exist between all objects, play a crucial role in understanding scenes. By addressing this gap, our method complements traditional approaches and significantly enhances model performance in SGG tasks.

2.2. Depth Estimation and 3D Reconstruction

Monocular depth estimation (MDE) has evolved significantly from early methods, which relied on handcrafted features and struggled with complex scenes due to their dependence on explicit depth cues [7, 17]. The introduction of deep learning transformed MDE, with Eigen *et al*. [15] pioneering a multi-scale fusion network for depth regression. Subsequent studies reframed regression as a classification task [4, 29], enhancing accuracy through improved priors and objective functions [23, 45]. For 3D reconstruction, classical methods computed dense depth per view [43] and used techniques like Delaunay triangulation [22] and Poisson surface reconstruction [21]. Recent deep learning approaches, such as ATLAS [37], NeuralRecon [48], and TransformerFusion [5], bypass traditional depth estimation by backprojecting 2D features into 3D space, though they incur high computational costs. These advancements have significantly improved the reconstruction of spatial scenes and the generation of inter-object spatial relationships.

2.3. Multimodal Large Language Models (MLLM)

In recent years, significant breakthroughs have been made in visual scene understanding. Models trained on large-scale image-text pairs [42] have demonstrated powerful performance in various vision tasks. Researchers have further enhanced the model's performance [25, 58], enabling Vision Language Models (VLM) to be applied in an expanding array of fields. Recently, the remarkable capabilities of Large Language Models (LLM) have led to a proliferation of LLM-based multimodal models [6, 11, 12, 24, 34, 62]. These MLLM inherit robust understanding and reasoning abilities. Additionally, designing prompts [6, 19, 53] enables MLLMs to reason about previously unseen tasks, and adding extra information [8, 18, 51] to inputs can enhance the MLLM's performance on tasks. We attempt to leverage this capability of MLLMs and have collected a dataset that can be utilized for these models, which we call the SpaceSGG dataset. This format is intended to enhance the model's understanding of both scenes and space, integrating text generation, object localization, relationship understanding, and spatial comprehension.

3. Methodology

To equip the model with scene graph generation capabilities and enhance spatial relation recognition, we first introduce the data construction pipeline for building the SpaceSGG dataset, which integrates spatial and scene graph information. Building upon this, we then detail the training

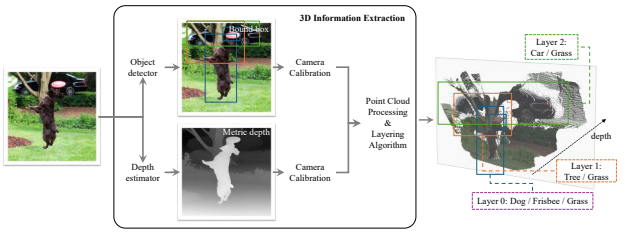


Figure 3. 3D Information Extraction: We retrieve the spatial layering distribution of the input images with the assistance of object detectors and depth estimator.

paradigm for LLaVA-SpaceSGG.

3.1. SpaceSGG Dataset Construction

We hypothesize that the model's inaccurate predictions of spatial relationships stem from a lack of relevant spatial information annotations and the inadequate integration of spatial and scene graph information. To address this, we design a two-stage data generation process, as shown in Figure 2. In the first stage, scene graph description (Section 3.1.1) and spatial layout are extracted from the 2D image (Section 3.1.2). In the second stage, the scene graph triplets and spatial layout are fused into a spatial scene graph description (Section 3.1.3), followed by spatial QA and spatial multi-turn conversations (Section 3.1.4). In this paper, we adopt the Llama 3 70B [2] as the data generator.

3.1.1 Scene Graph Description Generation

Scene graph description provides a detailed description, which serves as an intermediate state for the model to further generate a layered, comprehensive description of the image. To achieve this goal, we query GPT-4V [1] to generate responses that link the objects and predicates mentioned in the generated response to specific regions within the image, by following work [51].

3.1.2 Spatial Relations Extraction

To ensure that the spatial scene graph data contains more accurate spatial relationships, we need to reconstruct 3D spatial relationships from 2D images. We developed a pipeline that converts 2D images into depth maps and then into 3D point clouds. Firstly we apply the depth-anything [57] model as our base depth detector and align the relative depth with each detected object (see Figure 3). Then, a camera calibration [13] has been appended including two parts: 1) estimating intrinsic parameters to convert depth maps into 3D point clouds, and 2) ensuring that scene relationships are consistent. This allows us to create a rotation matrix to convert the point cloud for each object. The point cloud data are saved accordingly with the objects for further conversion.

Algorithm 1 Relative Spatial Relation of Object A and B

Input: object A, object B, image
depthmap ← DepthAnything(image)
pointcloud ← CameraCalibration(depthmap)
points A ← pointcloud[object A]
points B ← pointcloud[object B]
z_range_a ← [points A.Z_min, points A.Z_max]
z_range_b ← [points B.Z_min, points B.Z_max]
if z_range_a.min < z_range_b.min and z_range_a.max > z_range_b.max **then**
    z_range_a covers z_range_b
end if
record the relative spatial relation of objects A and B.

Algorithm 2 Devide objects into layers

Input: objects, image
Initialize layer_list.
for object A **in** objects **do**
    Initialize basic_layer_element_flag with True.
    for object B **in** objects **do**
        if object A is overlapping with object B **then**
            basic_layer_element_flag ← False
        end if
    end for
    if basic_layer_element_flag **then**
        Add object A into layer_list
    end if
end for
Sort objects in layer_list with depth
for object A **in** layer_list **do**
    Initialize sub_layer_list of A
    for object B **in** objects **do**
        if object A is covered by object B **then**
            Add object B into sub_layer_list of A
        end if
    end for
end for
record the layer_list and corresponding sub_layer_list.

3.1.3 SpaceSGG Description Generation

After stage 1, we focus on naturally integrating spatial layout and scene graph information in stage 2. We define explicit spatial layering to cluster and organize objects in space using the following algorithms (see Algorithm 1 and Algorithm 2): 1) one object with a depth range (by retrieving the minimum and maximum z-axis value) enclosed inside other objects' depth range is considered covered by others; 2) objects not covering any other objects are considered to be the basic objects representing an individual layer; 3) we check all the objects and select all the basic objects as the first elements in each layer; 4) we check all the

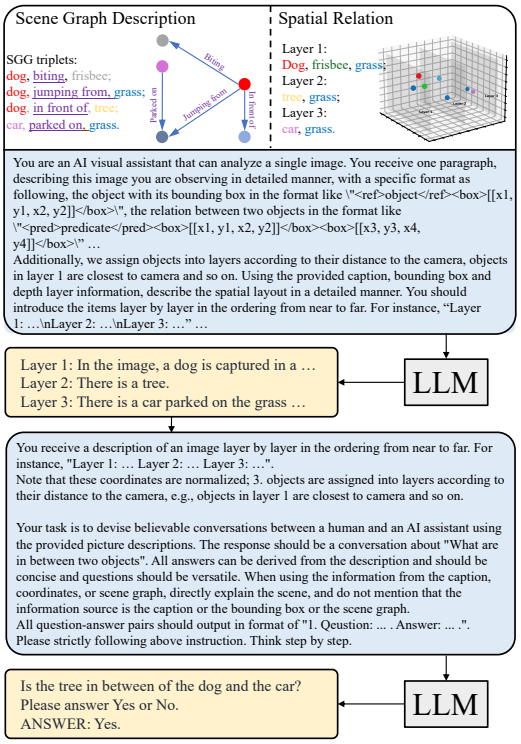


Figure 4. An example of SpaceSGG-Desc, SpaceSGG-QA, and SpaceSGG-Conv generation process.

other objects and assign them to different layers, such that larger objects may be added to multiple layers. This definition emphasizes the spatial information of objects for further processing. Then, we use a large language model [2] to reorganize the language from the previous step, emphasizing the spatial position information of objects and expressing the hierarchical information of objects in space. More specifically, we prepare the structural scene graph description (See Figure 4 top left) and the spatial relations after spatial layering (See Figure 4 top right) as input, and apply well-designed instructions (See Figure 4 blue regions) for prompting LLMs. The final output is grouped by layers termed **SpaceSGG-Desc**.

3.1.4 SpaceSGG QA and Conversation Generation

To enhance the model's spatial understanding and scene graph generation, we generate spatial question-answer pairs from scene graph descriptions, including single-turn QA (**SpaceSGG-QA**) and multi-turn conversation data (**SpaceSGG-Conv**) using LLM queries. Prompts (see Figure 4 bottom) are designed to split scene graph descriptions, triplets, and object depth distributions into questions about front-back judgment, up-down judgment, multi-object sorting, occlusion, and more (see Supplementary Materials A for examples). The SpaceSGG dataset combines spatial descriptions (SpaceSGG-Desc), single-turn QA (SpaceSGG-QA), and multi-turn conversations (SpaceSGG-Conv), encompassing 20K diverse scenes and their spatial structures.

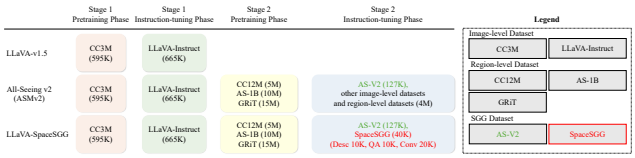


Figure 5. Our proposed training paradigm and used training dataset.

3.2. Training Paradigm

To enhance spatial SGG knowledge transfer to MLLMs, we propose a specialized training paradigm inspired by LLaVA-1.5 [34] and ASMv2 [51]. As shown in Figure 5, our approach consists of two stages, each with a pre-training and instruction-tuning phase. In the first stage, we align modality features using image-level datasets, following the LLaVA-1.5 [34] setup. In the second stage, pre-training uses region-level datasets to refine feature discovery and grounding, while instruction-tuning combines our proposed dataset with existing SGG data to enhance the model's understanding of scene relationships and spatial layouts. The datasets used in each stage are detailed below.

In the first stage, we use image-level datasets CC3M [46] and LLaVA-Instruct [34]. For the second stage, we use region-level datasets CC12M [46], AS-1B [52], and GRiT [39] to enhance the model's ability to discover subtle features and improve grounding. Unlike ASMv2, which incorporates 4M images from additional datasets like OCR-VQA [36] and TextVQA [47], our instruction-tuning phase focuses solely on SGG datasets, including AS-V2 [51] and the proposed SpaceSGG dataset.

4. Experiments

We conduct both quantitative and qualitative analysis with state-of-the-art methods on the public PSG dataset [55] and the proposed spatial relation validation set. Then, we design several ablation study settings to show the effectiveness of our training paradigm and SpaceSGG dataset.

4.1. Validation Sets and Metrics

PSG Validation Set. We use the PSG dataset [55] to evaluate the open-vocabulary SGG capabilities of existing models. The dataset includes 49K images and 56 relationships, six of which are positional (e.g., over, in front of, beside, on, in, attached to). We evaluate both closed-set methods [31, 49, 54, 55, 59] and open-set methods [51, 61].

Spatial Relation Validation Set. To assess spatial understanding, we use the proposed spatial relation validation

Model	Recall	mRecall
Close-ended SGG		
IMP	16.5	6.5
MOTIFS	20.0	9.1
VCTree	20.6	9.7
GPSNet	17.8	7.0
PSGFormer	18.6	16.7
Open-ended SGG		
TextPSG	4.8	–
ASMv2	14.2	10.3
LLaVA-SpaceSGG	15.43	13.23

Table 1. Open-vocabulary SGG performance comparison between our model and other specialist models. The red denotes the best results across all methods.

set. We randomly select 30 images from COCO-Val-2017 and generate two types of questions (QA and multi-turn conversations) using the data generation pipeline in Section 3.1.4. These are manually annotated as single-choice QA with factual corrections, resulting in 271 questions. This benchmark evaluates the model's spatial understanding of scenes.

Metrics. Following [51, 55, 61], we report triplet Recall and mean Recall (mRecall) for each predicate category in the open-vocabulary SGG task. A scene graph consists of triplets (subject, predicate, object), and a triplet is considered correct if the phrase labels are accurate and the subject and object locations match the ground truth with an Intersection over Union (IoU) greater than 0.5. Recall and mRecall are then computed as follows:

$$\text{Recall} = \frac{\text{Number of predicates}}{\text{Total number of ground truth relationships}}. \quad (1)$$

$$\text{mRecall} = \frac{1}{N}\sum_{i=1}^{N} \text{Recall}\ (i \in \textit{relation classes}). \quad (2)$$

4.2. Comparison with SoTA Methods

To test the effectiveness of our SpaceSGG dataset and training method, we compared LLaVA-SpaceSGG with state-of-the-art models [31, 49, 51, 54, 55, 59, 61] on the PSG [55] dataset and our proposed spatial relation validation set.

4.2.1 PSG Validation Set

We compare our model on the PSG validation set under the open-vocabulary SGG setting against open-set models [51, 61] and closed-set models [31, 49, 54, 55, 59]. As shown in Table 1, our model achieves state-of-the-art performance, outperforming ASMv2 by 8.6% in Recall and 28.4% in mRecall. Against closed-set models, it demonstrates strong performance with a recall of 15.43 and a mean recall of 13.23.

As shown in Figure 6 (bottom example), TextPSG often generates redundant relationships. However, our model produces concise scene graphs with accurate spatial relations (e.g., "in front of," "inside of," "beside"), leading to higher recall. Compared to ASMv2, our model captures more nuanced object relationships, enriching scene graphs and improving scene understanding. Additional results are provided in Supplementary Materials B.

Model	Accuracy (%)
Random Choice	25.00
LLaVA-1.5-13B	45.13
ASMv2-13B	50.52
LLaVA-SpaceSGG	52.48

Table 2. Comparison of model accuracies for spatial understanding tasks. Our model outperforms established benchmarks.

4.2.2 Spatial Relation Validation Set

To validate the spatial capabilities of LLaVA-SpaceSGG, we compare its spatial relationship prediction accuracy with state-of-the-art open-source MLLMs [33, 51] on our spatial relation validation set. As shown in Table 2, our model outperforms existing methods, with ASMv2-13B [52] being the closest competitor due to its use of scene graph data, though it struggles with 3D spatial contexts.

Figure 6 (top example) illustrates that, compared to TextPSG and ASMv2, our method captures more detailed spatial relationships, such as front-back orientations (blue) and distinctions between terms like "on the left," "beside," "attached to," "on," and "over" (purple). This demonstrates the value of integrating high-quality spatial and scene graph data, which current state-of-the-art models lack.

4.3. Ablation Studies

In this section, we conduct ablation studies to prove the effectiveness of our training paradigm and our proposed SpaceSGG dataset. All ablation settings can be found in Figure 7. We first compare the effectiveness of the training paradigm in three settings (see Figure 7 ablations on training paradigms). Then, we conduct five ablation studies to validate the effectiveness of our proposed SpaceSGG dataset (see Figure 7 ablations on data combinations).

4.3.1 Comparison on Different Training Paradigms

We conduct experiments under three training paradigms to evaluate the interaction of our SpaceSGG dataset with other

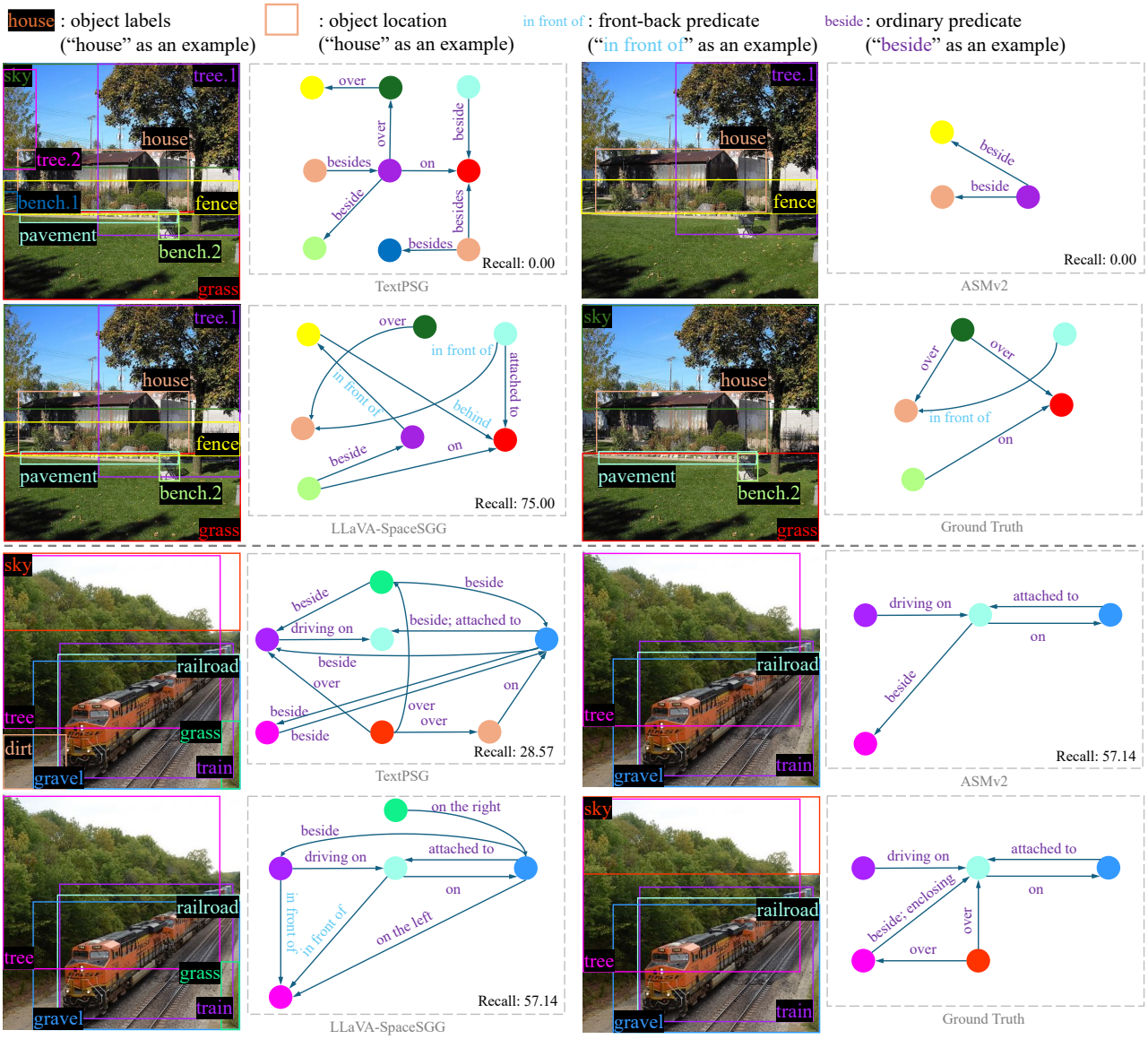


Figure 6. Qualitative result of open-vocabulary SGG, particularly from traditionally state-of-arts models. Note that the cyan-coloured predicate denotes a precise front-back relationship.

data, as shown in Table 3. The LLaVA-SpaceSGG-ab-train-1 setting performs worse than LLaVA-SpaceSGG, with a 6% drop in recall, 21.9% in mean recall, and reduced spatial validation accuracy, highlighting the benefits of mixing SpaceSGG with the AS-V2 dataset during stage 2 instruction tuning. LLaVA-SpaceSGG-ab-train-2, based on an additional SFT phase, also underperforms, demonstrating the efficiency and effectiveness of our two-stage training paradigm. Lastly, LLaVA-SpaceSGG-ab-train-3, which uses the existing ASMv2 training approach not tailored for SGG, achieves lower performance, confirming the superiority of our training design for SGG tasks.

4.3.2 Comparison on Different Data Combinations

To validate the effectiveness of the proposed SpaceSGG dataset, we conducted an ablation study by excluding specific terms from the SpaceSGG dataset to create five settings. First, LLaVA-SpaceSGG-ab-train-1, trained without the SpaceSGG dataset, shows reduced performance on the PSG validation set and a significant drop in the spatial benchmark, emphasizing the dataset's importance. Second, LLaVA-SpaceSGG-ab-data-1, LLaVA-SpaceSGG-ab-data-2, and LLaVA-SpaceSGG-ab-data-3 perform worse than the full model, as shown in Table 4, underscoring the value of data combination. Third,

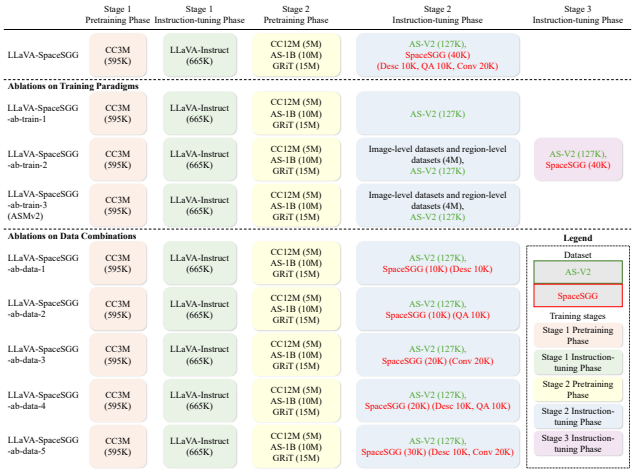


Figure 7. We conduct two types of ablation studies: training paradigm and data combination.

Ablation Setting	Recall	mRecall	Accuracy (%)
LLaVA-SpaceSGG -ab-train-1	14.41	10.32	1.47
LLaVA-SpaceSGG -ab-train-2	13.97	10.2	50.52
LLaVA-SpaceSGG -ab-train-3	14.2	10.3	45.7
LLaVA-SpaceSGG	15.43	13.23	52.48

Table 3. We conducted ablation studies based on different training stages. For the specific training set-ups of various settings, please refer to Figure 7.

LLaVA-SpaceSGG-ab-data-4 and LLaVA-SpaceSGG-ab-data-5, which include SpaceSGG-QA and SpaceSGG-Conv alongside SpaceSGG-Desc, achieve higher spatial benchmark accuracy but lower Recall and mRecall than the baseline, which better balances these metrics.

4.3.3 Comparison on Different Generative Models in Pipeline

To evaluate the role of data generator in the proposed pipeline, we replaced the default Llama 3 70B [2] with alternatives, including Qwen2.5 72B [41] and GPT-4o [38]. The results are shown in Table 5, which demonstrate that the data quality remains consistent across different generative models, with performance on Open Vocabulary SGG and the Spatial Validation set varying minimally compared to the final chosen model (i.e., Llama 3 70B). These findings suggest that the choice of generative model has a negligible impact on overall data quality.

Overall, these results demonstrate that the spatial and SGG information in our dataset is highly effective and significantly enhances the model's performance on SGG tasks.

Ablation Setting	**Recall**	**mRecall**	**Accuracy (%)**
LLaVA-SpaceSGG -ab-data-1	14.86	10.92	12.74
LLaVA-SpaceSGG -ab-data-2	14.53	11.07	37.21
LLaVA-SpaceSGG -ab-data-3	14.24	12.27	4.41
LLaVA-SpaceSGG -ab-data-4	14.39	11.26	53.39
LLaVA-SpaceSGG -ab-data-5	14.5	10.2	24.03
LLaVA-SpaceSGG	15.43	13.23	52.48

Table 4. We experimented with different mixing ratios of our generated data. The red, blue, and green colors denote the best, the second highest and the third highest results, respectively. For detailed experimental settings, please refer to Figure 7.

Ablation Setting	**Recall**	**mRecall**	**Accuracy (%)**
LLaVA-SpaceSGG -ab-Qwen2.5	14.22	9.53	51.68
LLaVA-SpaceSGG -ab-GPT-4o	13.99	10.94	53.725
LLaVA-SpaceSGG (Llama3)	15.43	13.23	52.48

Table 5. We experimented with different generated data by generative models.

5. Conclusions

In this paper, we tackle the problem of open-vocabulary scene graph generation by enhancing the spatial relations. Since most of the existing datasets often overlook 3D relations in SGG, we propose a data generation pipeline that integrates both 2D and 3D scene information to obtain more comprehensive relations. It results in 10K spatial scene detailed descriptions, 20K question answers, and 20K multi-turn conversations. Built upon it, we present the LLaVA-SpaceSGG model. Specifically, we explore a task-specific training paradigm, which contains two stages that improve the model's ability to perceive spatial relations in complex visual scenes. Finally, we compare the proposed LLaVA-SpaceSGG on a well-known PSG dataset and our proposed spatial relation validation set. The experiment results show that LLaVA-SpaceSGG surpasses the current state-of-the-art models in the open-vocabulary SGG task, achieving an 8.6% improvement in recall and a 28.4% improvement in mRecall. It also performs better than other MLLMs in producing spatial relations. In the future, we aim to further enhance the model's visual understanding and reasoning capabilities by using high-quality and diverse annotations.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 4

[2] AI@Meta. Llama 3 model card. 2024. 4, 5, 8

[3] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015. 1

[4] Shariq Farooq Bhat, Ibraheem Alhashim, and Peter Wonka. Adabins: Depth estimation using adaptive bins. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4009–4018, 2021. 3

[5] Aljaz Bozic, Pablo Palafox, Justus Thies, Angela Dai, and Matthias Nießner. Transformerfusion: Monocular rgb scene reconstruction using transformers. *Advances in Neural Information Processing Systems*, 34:1403–1414, 2021. 3

[6] Tom B Brown. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020. 3

[7] LIU Ce. Sift flow: Dense correspondence across different scenes. *ECCV 2008*, 2008. 3

[8] Boyuan Chen, Zhuo Xu, Sean Kirmani, Brian Ichter, Danny Driess, Pete Florence, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. *arXiv preprint arXiv:2401.12168*, 2024. 3

[9] Lin Chen, Jisong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. Sharegpt4v: Improving large multi-modal models with better captions. *arXiv preprint arXiv:2311.12793*, 2023. 1

[10] Tianshui Chen, Weihao Yu, Riquan Chen, and Liang Lin. Knowledge-embedded routing network for scene graph generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6163–6171, 2019. 3

[11] Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *arXiv preprint arXiv:2404.16821*, 2024. 3

[12] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24185–24198, 2024. 1, 3

[13] An-Chieh Cheng, Hongxu Yin, Yang Fu, Qiushan Guo, Ruihan Yang, Jan Kautz, Xiaolong Wang, and Sifei Liu. Spatialrgpt: Grounded spatial reasoning in vision language model. *arXiv preprint arXiv:2406.01584*, 2024. 4

[14] Arda Duzceker, Silvano Galliani, Christoph Vogel, Pablo Speciale, Mihai Dusmanu, and Marc Pollefeys. Deepvideomvs: Multi-view stereo on video with recurrent spatio-temporal fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15324–15333, June 2021. 2

[15] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. *Advances in neural information processing systems*, 27, 2014. 3

[16] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913, 2017. 1

[17] Derek Hoiem, Alexei A Efros, and Martial Hebert. Recovering surface layout from an image. *International Journal of Computer Vision*, 75:151–172, 2007. 3

[18] Shaohan Huang, Li Dong, Wenhui Wang, Yaru Hao, Saksham Singhal, Shuming Ma, Tengchao Lv, Lei Cui, Owais Khan Mohammed, Barun Patra, et al. Language is not all you need: Aligning perception with language models. *Advances in Neural Information Processing Systems*, 36, 2024. 3

[19] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *European Conference on Computer Vision*, pages 709–727. Springer, 2022. 3

[20] Justin Johnson, Ranjay Krishna, Michael Stark, Li-Jia Li, David Shamma, Michael Bernstein, and Li Fei-Fei. Image retrieval using scene graphs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3668–3678, 2015. 1

[21] Michael Kazhdan, Matthew Bolitho, and Hugues Hoppe. Poisson surface reconstruction. In *Proceedings of the fourth Eurographics symposium on Geometry processing*, volume 7, 2006. 3

[22] Der-Tsai Lee and Bruce J Schachter. Two algorithms for constructing a delaunay triangulation. *International Journal of Computer & Information Sciences*, 9(3):219–242, 1980. 3

[23] Bo Li, Chunhua Shen, Yuchao Dai, Anton Van Den Hengel, and Mingyi He. Depth and surface normal estimation from monocular images using regression on deep features and hierarchical crfs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1119–1127, 2015. 3

[24] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023. 3

[25] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705, 2021. 3

[26] Yikang Li, Wanli Ouyang, Bolei Zhou, Jianping Shi, Chao Zhang, and Xiaogang Wang. Factorizable net: an efficient subgraph-based framework for scene graph generation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 335–351, 2018. 1

[27] Yikang Li, Wanli Ouyang, Bolei Zhou, Kun Wang, and Xiaogang Wang. Scene graph generation from objects, phrases and region captions. In *Proceedings of the IEEE international conference on computer vision*, pages 1261–1270, 2017. 1, 3

[28] Yanda Li, Chi Zhang, Gang Yu, Zhibin Wang, Bin Fu, Guosheng Lin, Chunhua Shen, Ling Chen, and Yunchao Wei. Stablellava: Enhanced visual instruction tuning with synthesized image-dialogue data. *arXiv preprint arXiv:2308.10253*, 2023. 1

[29] Zhenyu Li, Xuyang Wang, Xianming Liu, and Junjun Jiang. Binsformer: Revisiting adaptive bins for monocular depth estimation. *IEEE Transactions on Image Processing*, 2024. 3

[30] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 1, 2

[31] Xin Lin, Changxing Ding, Jinquan Zeng, and Dacheng Tao. Gps-net: Graph property sensing network for scene graph generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3746–3753, 2020. 5, 6

[32] Fangyu Liu, Guy Emerson, and Nigel Collier. Visual spatial reasoning. *Transactions of the Association for Computational Linguistics*, 11:635–651, 2023. 1

[33] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306, 2024. 1, 6

[34] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024. 3, 5

[35] Cewu Lu, Ranjay Krishna, Michael Bernstein, and Li Fei-Fei. Visual relationship detection with language priors. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, pages 852–869. Springer, 2016. 1, 2

[36] Anand Mishra, Shashank Shekhar, Ajeet Kumar Singh, and Anirban Chakraborty. Ocr-vqa: Visual question answering by reading text in images. In *ICDAR*, 2019. 5

[37] Zak Murez, Tarrence Van As, James Bartolozzi, Ayan Sinha, Vijay Badrinarayanan, and Andrew Rabinovich. Atlas: End-to-end 3d scene reconstruction from posed images. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VII 16*, pages 414–431. Springer, 2020. 3

[38] OpenAI. Gpt-4o system card, August 2024. 8

[39] Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. Kosmos-2: Grounding multimodal large language models to the world. *arXiv preprint arXiv:2306.14824*, 2023. 5

[40] Tao Pu, Tianshui Chen, Hefeng Wu, Yongyi Lu, and Liang Lin. Spatial-temporal knowledge-embedded transformer for video scene graph generation. *IEEE Transactions on Image Processing*, 2023. 1

[41] Qwen Team. Qwen2.5: A party of foundation models, September 2024. 8

[42] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 1, 3

[43] Johannes L Schönberger, Enliang Zheng, Jan-Michael Frahm, and Marc Pollefeys. Pixelwise view selection for unstructured multi-view stereo. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III 14*, pages 501–518. Springer, 2016. 3

[44] Brigit Schroeder and Subarna Tripathi. Structured query-based image retrieval using scene graphs. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 178–179, 2020. 1

[45] Shuwei Shao, Zhongcai Pei, Weihai Chen, Xingming Wu, and Zhengguo Li. Nddepth: Normal-distance assisted monocular depth estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7931–7940, 2023. 3

[46] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, 2018. 1, 5

[47] Amanpreet Singh, Vivek Natarjan, Meet Shah, Yu Jiang, Xinlei Chen, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8317–8326, 2019. 5

[48] Jiaming Sun, Yiming Xie, Linghao Chen, Xiaowei Zhou, and Hujun Bao. Neuralrecon: Real-time coherent 3d reconstruction from monocular video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15598–15607, 2021. 3

[49] Kaihua Tang, Hanwang Zhang, Baoyuan Wu, Wenhan Luo, and Wei Liu. Learning to compose dynamic tree structures for visual contexts. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6619–6628, 2019. 5, 6

[50] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023. 2

[51] Weiyun Wang, Yiming Ren, Haowen Luo, Tiantong Li, Chenxiang Yan, Zhe Chen, Wenhai Wang, Qingyun Li, Lewei Lu, Xizhou Zhu, et al. The all-seeing project v2: Towards general relation comprehension of the open world. *arXiv preprint arXiv:2402.19474*, 2024. 1, 2, 3, 4, 5, 6

[52] Weiyun Wang, Min Shi, Qingyun Li, Wenhai Wang, Zhenhang Huang, Linjie Xing, Zhe Chen, Hao Li, Xizhou Zhu, Zhiguo Cao, et al. The all-seeing project: Towards panop-

tic visual recognition and understanding of the open world. *arXiv preprint arXiv:2308.01907*, 2023. 5, 6

[53] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022. 2, 3

[54] Danfei Xu, Yuke Zhu, Christopher B Choy, and Li Fei-Fei. Scene graph generation by iterative message passing. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5410–5419, 2017. 3, 5, 6

[55] Jingkang Yang, Yi Zhe Ang, Zujin Guo, Kaiyang Zhou, Wayne Zhang, and Ziwei Liu. Panoptic scene graph generation. In *European Conference on Computer Vision*, pages 178–196. Springer, 2022. 2, 3, 5, 6

[56] Jianwei Yang, Jiasen Lu, Stefan Lee, Dhruv Batra, and Devi Parikh. Graph r-cnn for scene graph generation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 670–685, 2018. 3

[57] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10371–10381, 2024. 2, 4

[58] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*, 2022. 1, 3

[59] Rowan Zellers, Mark Yatskar, Sam Thomson, and Yejin Choi. Neural motifs: Scene graph parsing with global context. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5831–5840, 2018. 5, 6

[60] Pengpeng Zeng, Lianli Gao, Xinyu Lyu, Shuaiqi Jing, and Jingkuan Song. Conceptual and syntactical cross-modal alignment with cross-level consistency for image-text matching. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 2205–2213, 2021. 1

[61] Chengyang Zhao, Yikang Shen, Zhenfang Chen, Mingyu Ding, and Chuang Gan. Textpsg: Panoptic scene graph generation from textual descriptions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2839–2850, 2023. 5, 6

[62] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023. 3