Synthetic Visual Genome

Jae Sung Park[1], **Zixian Ma**[1], **Linjie Li**[1], **Chenhao Zheng**[1], **Cheng-Yu Hsieh**[1],
Ximing Lu[1], **Khyathi Chandu**[2], **Quan Kong**[4],
Norimasa Kobori[4], **Ali Farhadi**[1,2], **Yejin Choi**[3], **Ranjay Krishna**[1,2]

[1]University of Washington, [2]Allen Institute for Artificial Intelligence,
[3]Stanford University, [4]Woven by Toyota

Abstract

Reasoning over visual relationships—spatial, functional, interactional, social, etc.—is considered to be a fundamental component of human cognition. Yet, despite the major advances in visual comprehension in multimodal language models (MLMs), precise reasoning over relationships and their generations remains a challenge. We introduce ROBIN*: an MLM instruction-tuned with densely annotated relationships capable of constructing high-quality dense scene graphs at scale. To train* ROBIN*, we curate* SVG[1]*, a synthetic scene graph dataset by completing the missing relations of selected objects in existing scene graphs using a teacher MLM and a carefully designed filtering process to ensure high-quality. To generate more accurate and rich scene graphs at scale for any image, we introduce* SG-EDIT*: a self-distillation framework where GPT-4o further refines* ROBIN*'s predicted scene graphs by removing unlikely relations and/or suggesting relevant ones. In total, our dataset contains* $146K$ *images and* $5.6M$ *relationships for* $2.6M$ *objects. Results show that our* ROBIN*-3B model, despite being trained on less than* 3 *million instances, outperforms similar-size models trained on over* 300 *million instances on relationship understanding benchmarks, and even surpasses larger models up to 13B parameters. Notably, it achieves state-of-the-art performance in referring expression comprehension with a score of* 88.2*, surpassing the previous best of* 87.4*. Our results suggest that training on the refined scene graph data is crucial to maintaining high performance across diverse visual reasoning tasks*[2]*.*

1. Introduction

Scholars have argued for decades that visual relationship reasoning is a fundamental characteristic of human intelligence [4, 5]. By reasoning over relationships, people can make sense of new scenes by stitching together individual objects and their pairwise relationships [6, 14, 26, 28]. For instance, relationships allow people to describe the photo in Figure 2 using a set of spatial (*woman-above-raft*), social (*child-cared_by-woman*), functional (*man-using-paddle*), interactional (*woman-holding-baby*) and emotional (*man-sharing_an_experience_with-family*) relationships. For a while now, multimodal language model (MLM) research has sought to develop models that can similarly comprehend such relationships [20, 29, 37, 51]. Yet despite all the progress, frontier MLMs still struggle to accurately express relationships; open-source MLMs perform even worse [75].

Instruction-tuning has been established as a successful mechanism to instill specific reasoning capabilities in MLMs [15]. However, instruction tuning an open-sourced MLM to understand diverse relationships is not possible today due to the lack of large-scale datasets focused on relationship understanding; existing efforts are limited to spatial reasoning derived from synthetic data [8]. Scene graphs could serve as potential sources for instruction tuning because they provide direct annotations of objects and their re-

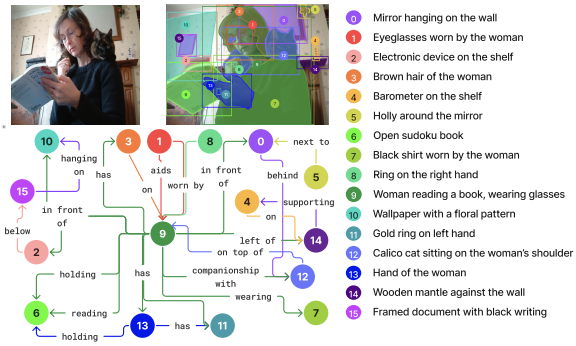

Figure 1. Example outputs of the Synthetic Visual Genome (SVG) dataset, the first automatically generated large-scale scene graph dataset, featuring diverse, open-set categories, and dense relationship annotations. On average, SVG has four times as many relations per object compared to Visual Genome [37], offering substantially richer relationship density than existing human annotations.

[1]Synthetic Visual Genome

[2]The SVG data, model checkpoints, and code are available at https://synthetic-visual-genome.github.io/

Figure 2. We build SVG dataset by leveraging GPT4V to complete missing relationships for selected objects spanning five categories: spatial, interactional, emotional, functional, and social. We then distill our ROBIN on these newly enriched annotations and use the trained model to generate complete, dense scene graphs.

lationships within scenes. Nevertheless, existing scene graph datasets have limitations in image coverage and relationship diversity. Visual Genome [37], for instance, contains a large number of spatial relationships but is limited in interactional, emotional, or functional relationships (Figure 1) and provides annotations for only 100K images. Additionally, its annotations are rather *sparse*, with only 1.5 relations annotated on average for every subject, and fails to enumerate all possible relationships in the scene, including some spatial ones like *woman, in front of, baby*. This is because even for humans, annotating every relationships for all objects is a cumbersome process, and scaling such detailed labeling efforts to address these issues will be impractical due to the immense time and cost required.

To overcome this bottleneck, we propose automating the process of generating densely annotated scene graphs. One option is to prompt frontier models like GPT-4V to generate instruction-tuning data, which has shown promise for many language and reasoning tasks [79]. However, such methods have found limited utility in computer vision as even current frontier models struggle to understand 3D structure and interactions between objects [17]. As a result, when prompted to generate scene graph data from scratch, they hallucinate and produce low-quality data [78].

In this work, we introduce a more reliable dataset bootstrapping framework, that allows us to build high quality scene-graph dataset at scale using limited amount of seed supervision data and frontier models. In particular, 1) we start out by leveraging a frontier model, GPT-4V, to enrich existing scene-graph datasets by completing the missing relationships based on the initial human annotations available instead of completely from scratch. This leads to a synthetic yet high-quality and dense scene graph dataset **SVG**, with 146K images annotated with 5.6M relationships for 2.6M objects. 2) Using SVG, we then train **Robin-3B**, a dedicated MLM that is able to generate dense scene graphs with enhanced relationship understanding and grounded reasoning. 3) Finally, we show how Robin-3B can facilitate a scalable synthetic scene-graph data generation pipeline, **SG-EDIT**, which uses Robin-3B to efficiently generate dense scene-graph data from scratch coupled with GPT-4V for further refinement.

Our experiments validate our contributions in three-fold. First, we validate the quality of SVG, by showing that training with SVG leads to an effective MLM capable of high-quality dense scene-graph generation. Second, we validate that SG-EDIT is an effective scene-graph dataset scaling solution, wherein the resultant enhanced Robin-3B, despite being trained on less than 3M instances, demonstrates superior performances than similar-size models trained on over 300 million instances and even larger models (up to 13B parameters) on various benchmarks in relationship understanding and grounded reasoning. Finally, we show that Robin-3B is capable of generating high-quality dense scene graphs, achieving the state of the art in open-ended scene graph generation.

2. Synthetic Visual Genome Data Pipeline

As existing scene graph datasets lack dense and diverse relationship annotations, we build a data pipeline that generates dense scene graphs at scale. Here, we make use of powerful visual reasoning capabilities from proprietary multimodal models (e.g., GPT-4) to systematically infer missing object relationships. The data curation involves two stages: 1) filling unannotated relationships for selected objects from existing image annotations (Sec 2.1), and 2) refining automatically generated dense scene graphs with GPT-4 editing (Sec 2.2). Comparison of Synthetic Visual Genome to existing scene graph dataset is shown in Table 1.

2.1. Relationship Completion on Seed Images

While it is tempting to rely on proprietary multimodal models to generate scene graphs from scratch, their limited grounding capabilities frequently yield hallucinations, making them unsuitable for accurate dataset curation (see Appendix A.2). Instead, we incorporate existing annotations to ensure correctness. Figure 2 illustrates the pipeline's first stage, where we construct scene graphs from seed images with dense, high-quality annotation. Specifically, we select a subset of COCO [44] images that include: a) object detection labels from COCO [44] and LVIS [21], b) region descriptions from RefCOCO [90] and Visual Genome (VG) [37], c) scene graphs from VG and GQA [25], and d) depth maps generated by the Depth-Anything model [86]. This yields 33K seed images with comprehensive annotations for each region.

Dataset	Images	Annotator	Region	Objects per image	Triplets per image	Predicates per region
VG [37]	108K	Human	Box	35.2	21.4	0.6
GQA [25]	85K	Human	Box	16.4	50.6	3.1
Open Images []	568K	Human	Box	8.4	5.6	0.7
PSG [84]	49K	Human	Seg + Box	11.2	5.7	0.6
SVG-RELATIONS	33K	GPT-4V	Seg + Box	13.2	25.5	1.9
SVG-FULL	113K	ROBIN + GPT-4o	Seg + Box	19.8	42.3	2.4

Table 1. Comparison among existing scene graph datasets and Synthetic Visual Genome.

Selecting semantically significant objects Next, we filter out object regions of low semantic significance to retain visually distinct and meaningful elements. Specifically, we use the Segment Anything Model (SAM) [36] and Semantic-SAM [40] to generate segmentation masks representing prominent objects and regions in each image. For each annotated region in our seed dataset, we compute the Intersection over Union (IoU) score between our annotated regions and the segmentation masks produced by SAM and Semantic-SAM. We then keep annotated regions with an IoU score greater than 0.5 with any of the segmentation masks.

Generating relationships with GPT-4V Building on the dense region annotations, we prompt GPT4-V to identify at lest $K = 5$ subjects in each scene. For each subject, we request 1) its description, and 2) a comprehensive list of its relationships with other objects, categorized into five types: spatial, interactional, functional, social, and emotional (see appendix Table 12 for prompt).

Filtering irrelevant relationships Despite using human annotations for object regions, GPT4-V is prone to generating relationships with errors. To enhance data reliability, we must implement a robust filtering strategy to remove such inconsistencies. Inspired by prior works [16, 60], we employ both rule-based and model-based approaches for data filtering. Specifically, we apply rule-based filtering to spatial relationships and model-based VQA filtering to the rest. In Appendix E.2, we elaborate on these filters and their impact on relationship distribution. We refer to this final filtered dataset with dense relationships as **SVG-RELATIONS**.

2.2. Dense Scene Graphs with GPT-4 Refinement

Having curated SVG-Stage1, we train a student model, referred to as ROBIN[3], to produce dense relationships for each object of interest. This way, we can now call ROBIN per object to produce relationships grounded in all provided regions, and thus generate complete, dense scene graphs for every image. Although densely annotated, SVG-RELATIONS is limited in coverage as it is solely curated from COCO [44]. As a result, the student model is likely to produce noisy scene graphs for images *in the wild*.

To address these challenges, we introduce a distillation pipeline, **SG-EDIT**, inspired by the iterative data improvement approach of Segment Anything [36]. Note GPT-4 still offers strong visual reasoning skills across images in the wild, making it an ideal automated editor for noisy object annotations. Rather than relying on costly human corrections that are difficult to scale, we thus use GPT-4[4] to edit Robin 's generated scene graphs by removing noisy relations, adding relevant ones, and specifying more precise object attributes. This yields a more accurate, refined set of dense scene graphs, on which we then retrain our model to further improve its performance. As shown in Figure 3, GPT4 can effectively refine scene graphs: for instance, removing an irrelevant "wearing" relation with a chair, expanding the person's description to "wearing maroon sweater," and introducing relational tags such as "posing for [13]" or "sitting on [1]."

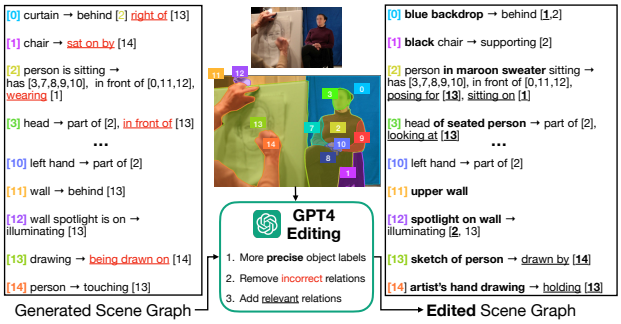

Figure 3. Using our SG-EDIT, GPT-4 can effectively provide more precise object descriptions with attributes (bolded), remove incorrect relations (highlighted in red), and add relevant (underlined).

We extend this pipeline to broaden the diversity of training data. Specifically, we bootstrap from available region, object, and relationship annotations, and have ROBIN generate dense relationships for each objet. GPT-4 then edits these scene graphs to ensure high quality. We apply this procedure to the ADE20K [101], PSG [84], and VG [37] datasets, generating 113K new synthetic scene graphs in total (25K/ 35K /53K respectively), which we refer as **SVG-FULL**.[5]

[3] Relation-Object Instruction Tuned Model

[4] gpt-4o-2024-08-06

[5] While we generate 93K scene graphs here, the procedure is scalable. Further exploration of scaling laws is left for future work."

3. ROBIN-3B

Using the SVG dataset, we present ROBIN-3B, an MLM trained to accurately reason about regions and produce dense scene graphs for any objects of interest. Specifically, ROBIN refers each region by its pixel-level segmentation masks [92] and bounding box coordinates in text [11, 52, 78].

3.1. Model architecture

The model architecture is similar to Osprey-7B [92], which consists of 1) a vision encoder that encodes the entire image to image tokens, 2) a pixel-level mask-aware extractor that embeds each segmentation to mask tokens, and 3) a language model (LM) that takes in image, mask, and text tokens to support any visual instruction and grounding tasks in the text space. We use ConvNext-Large [27] as visual and mask encoder, and Qwen2.5-3B as our LM [83]. Vision and mask projection layers are added to project the vision and mask tokens into text embeddings that will be passed to the LM (see appendix Figure 10 for model architecture and details).

3.2. Training stages

We now describe three training stages that progressively distill scene-graph reasoning into the model.

Stage 0: Image-segmentation-text alignment In this initial stage, we adapt the Qwen2.5-3B LM to process both image and mask tokens, following the three phases in [92]. The vision encoders are kept frozen, whereas the mask encoders, projection layers and the language model are trained. We use LLaVA-Pretrain-558K image-text pairs [95], and Osprey-724K instruction dataset [92], totaling 1.28M instances.

Stage 1: Instruction tuning with scene graphs Next, we unfreeze the visual encoder and train ROBIN-3B on three core data categories: (1) Visual Instructions, (2) Grounding, and (3) Scene Graphs. The scene graph data includes PSG [84], Visual Genome (VG)[37], and our own SVG-Stage1, yielding 1.98M total examples. A complete breakdown of these datasets, along with data formatting and distribution details, is presented in Appendix C and Table 7. We refer to the resulting model trained with this mixture as Robin-3B (Stage 1).

Stage 2: Distillation with GPT4 edited scene graphs Finally, as described in Sec 2.2, we use Robin (Stage 1) to generate complete scene graphs and apply our SG-EDIT pipeline with GPT-4 to build SVG-FULL. We then replace SVG-RELATIONS with SVG-FULL, remove the OCR-focused portion of the visual-instruction data, and continue training Robin on these updated inputs, resulting in our final Robin-3B. Details of this data mixture are shown in Table 8.

4. Experiments

In our experiments, we compare ROBIN-3B with the state-of-the-art (SoTA) open-source MLMs across a set of visual reasoning tasks focused on relationship understanding. To see the performance gain from SG-EDIT, we separately evaluate ROBIN-3B trained with Stage 1 and compare it against our ROBIN-3B model with Stage 2 training. Additionally, we assess whether incorporating scene graph training enhances the model's capabilities in grounding and region understanding. Lastly, we directly evaluate ROBIN-3B on scene graph generation to show its ability to produce accurate scene graphs. Prompts used for each task and dataset are provided in Appendix D.

4.1. Relationship understanding benchmarks

We first run evaluations on visual question answering benchmarks that focus on relationship understanding such as: GQA [25] that was originally designed for relational understanding using image scene graphs, MMBench [50], SeedBench [38], Visual Spatial Reasoning (VSR) [46], CRPE [53], SugarCrepe [24], and What's Up [31]. MMBench and SeedBench are included for their coverage of spatial relationship and object interaction understanding. The CPRE dataset focuses on relation comprehension that also includes abnormal relations generated with synthetic images. We evaluate on the subject-predicate-object split targeted for relation comprehension, and report the overall average, denoted as Relation. In SugarCrepe, we use the "replace-relation" split that and follow their approach of framing the evaluation as a binary multiple choice problem suitable for MLMs. In What's Up, we evaluate on the 820 images consisting of unambiguous object-relationships capture in controlled environments (denoted as "Controlled").

We present the results in Table 2. Among the models of similar size ($\leq$4B parameters) [1, 7, 43, 82, 87, 96], we observe that Robin-3B shows the strongest performance on MMBench, SugarCrepe, and Whats Up benchmarks, and competitive performance on others. Notably, despite being trained on fewer than 3M instances, Robin-3B surpasses Phi-3-Vision and BLIP-3 on the What's Up dataset by a significant margin (80.1% vs. 78.7% and 78.2%, respectively). This is particularly impressive given that Phi-3-Vision and BLIP-3 have undergone extensive pre-training on up to 300 million instances. In VSR, while Robin-3B seems to fall behind the BLIP-3 models, this is mostly because the dataset was seen in training stages of BLIP-3; on the other hand, our model greatly exceeds Phi-3-Vision (70.5% vs. 67.8%) when evaluated in a zero-shot manner, showcasing our model's strengths in understanding spatial relations in general[6].

[6]This explains the differing results between Phi-3-Vision and BLIP-3, which despite using the same LMs, show conflicting performance on VSR and What's Up – both of which assess spatial relation understanding.

Model	LLM	GQA [25]	MMBench [50] Dev-EN	SEED [38] Image	VSR [46] ZS-test	CRPE [78] Relation	SugarCrepe [24] Relation	What's Up? [32] Controlled
$\leq$ 4B Models								
VILA1.5-3B [43]	ShearedLLaMA-2.7B	61.5*	63.4	67.9	61.0	67.8	86.3	50.6
MiniCPM-V2.0-3B [87]	Mini-CPM-2.4B		69.7	67.1	68.2	68.1	86.6	54.8
InternVL2-2B [7]	InternLM-2B	61.0*	73.2	70.9	69.0*	65.8	85.5	74.4
Phi-3-Vision [1]	Phi-3-mini-4B	-	74.2	71.0	67.8	71.6	88.7	78.7
BLIP-3-single-image [82]	Phi-3-mini-4B	-	76.0	71.8	72.5*	72.4	89.0	78.2
BLIP-3-interleave [82]	Phi-3-mini-4B	-	76.8	72.2	**72.6***	**72.5**	88.3	76.3
MM1-3B [58]	MM1-3B	-	67.8	68.8	-	-	-	-
MM1.5-3B [96]	MM1-3B	-	-	**72.4**	-	-	-	-
Robin-3B (Stage 1)	Qwen2.5-3B	60.8*	**77.1**	69.4	69.7	64.8	87.9	76.7
Robin-3B	Qwen2.5-3B	61.6*	77.0	70.6	70.5	68.0	**89.4**	**80.1**
7B - 13B Models								
InstructBLIP [15]	Vicuna-7B	49.2	36.0	-	54.3	-	-	-
Shikra [11]	Vicuna-7B	-	58.8	-	63.3	-	-	-
QwenVL-7B [3]	Qwen-7B	59.3*	38.2	62.3	63.8	38.5	77.9	35.5
Mini-GPT-v2 [10]	LLaMA-2-7B	60.3*	-	-	62.9	-	-	-
LLaVA-1.5-7B [95]	Vicuna1.5-7B	62.0*	65.2	66.1	63.4	55.6	84.6	39.5
LLaVA-1.5-13B [95]	Vicuna1.5-13B	64.2*	67.4	70.2	65.3	66.9	87.1	49.5
ShareGPT4V-7B [12]	Vicuna1.5-7B	-	68.8	69.7	63.1	60.6	83.5	52.1
ShareGPT4V-13B [12]	Vicuna1.5-13B	-	-	70.8	67.5	55.0	85.1	55.7
LLaVA-1.6-Next-7B [48]	Vicuna1.5-7B	64.2*	67.4	70.2	68.6	69.6	85.7	43.3
LLaVA-1.6-Next-13B [48]	Vicuna1.5-13B	65.4*	70.0	71.9	65.8	69.9	87.8	56.8
VisionLLM v2-Chat [80]	Vicuna1.5-7B	**65.1***	**77.1**	71.7	-	-	-	-
ASMv2-13B [78]	Vicuna1.5-13B	63.9*	74.4	65.0	69.5	64.5	87.5	54.0

Table 2. Relationship understanding performance comparison between Robin-3B and state of the art multimodal language models, grouped by parameter size ($\leq$ 4B and 7B–13B). Bold numbers indicate the best performance across the models. (*): data was included in training.

Next, we compare between the Stage 1 and Stage 2 trained models. While there is a negligible performance decrease on MMBench, our Stage 2 model shows improvements over Stage 1 model in almost every benchmark. Notably, we see substantial gains in CRPE (68.0% vs 64.8%), SugarCrepe (89.4% vs 87.9%), and Whats Up (80.1% vs 76.7%).

When compared with larger models, ROBIN demonstrates superior performance in in VSR, SugarCrepe, and What's Up This reveals a lack of relationship understanding in current visual instruction frameworks and shows the benefits of incorporating scene graphs into the instruction tuning dataset to enhance performance in relationship understanding tasks. Additionally, we outperform ASMv2-13B [78] which is another MLLM specialized for object-relation understanding on all benchmarks except GQA. Notably, on their proposed CREPE benchmark, ROBIN surpasses them by a significant margin (68.0% vs. 64.5%), highlighting the effectiveness of our relationship training framework over theirs.

4.2. Referring expression comprehension

We next assess the grounded reasoning capabilities of our model on referring expression comprehension tasks using the RefCOCO, RefCOCO+, and RefCOCOg datasets [54, 90]. We prompt our models to provide a bounding box for each description, and report Recall@1 (IoU $>$ 0.5). As shown in Table 3, Robin-3B achieves the highest average accuracy overall, with an average of 88.2% in the test split, surpassing larger models such as ASM-V2-13B (87.3%). We significantly outperform MM1.5-3B, the previous state of the art among the 3B models (85.6%) that has been trained with more than 1M instances for grounding. The improvements from Stage 1 to Stage 2 (86.0% $\rightarrow$ 88.2%) again highlight the effectiveness of our self-distillation approach in enhancing grounded reasoning abilities. These results demonstrate that incorporating scene graph training enhances the grounding capabilities, even outperforming models with much larger parameter counts and pre-training data.

4.3. Region recognition

We evaluate our model on open-vocabulary region recognition tasks, namely semantic segmentation on the ADE20k dataset [101] and region classification on the LVIS and PACO datasets [21, 65]. For each task, the model generates a description or category label for the region specified by a segmentation mask and bounding box; we then convert text outputs to class labels using SentenceBERT [68] similarity, following [92].

Table 4 shows that ROBIN-3B outperforms 7B-scale models across all metrics. Against Osprey-7B, it achieves gains of +2.3% PQ, +4.0% mAP, and +2.3% mIoU on ADE20K, and exceeds Osprey-7B by +7.4% SS and +11.7% S-IOU on LVIS. Moreover, compared to VisionLLMv2-7B, our model improves LVIS scores by +5.3% SS and +7.2% S-IOU and PACO scores by +9.9% SS and +16.6% S-IOU. Finally, progressing from Stage1 to Stage2 yields an additional +2.6% SS and +3.8% S-IOU on PACO, pushing ROBIN-3B beyond Osprey-7B.

Model	RefCOCO [90]			RefCOCO+ [90]			RefCOCOg [54]		Avg	Avg_{test}
	Val	Test-A	Test-B	Val	Test-A	Test-B	Val	Test		
Shikra-7B [11]	87.0	90.6	80.2	81.6	87.4	72.1	82.3	82.2	82.9	82.5
MiniGPT-v2-7B [10]	88.1	91.3	84.3	79.6	85.5	73.3	84.2	84.3	83.8	83.7
QwenVL-7B [3]	88.6	92.3	84.5	82.8	88.6	76.8	86.0	86.3	85.7	85.7
Ferret-7B [89]	87.5	91.3	82.4	80.8	87.4	73.1	83.9	84.8	83.9	83.8
Groma-7B [52]	89.5	92.1	86.3	83.9	88.9	78.1	86.4	87.0	86.5	86.5
VisionLLMv2-Chat [80]	90.0	93.1	87.1	81.1	87.3	74.5	85.0	86.4	85.6	85.7
Ferret-13B [89]	89.5	92.4	84.4	82.8	88.1	75.2	85.8	86.3	85.6	85.3
ASM-V2-13B [78]	90.6	94.2	86.2	84.8	90.8	76.9	87.5	88.3	87.4	87.3
InternVL2-2B [7]	82.3	88.2	75.9	73.5	82.8	63.3	77.6	78.3	77.8	77.7
Phi-3-Vision-4B [1]	-	46.3	36.1	-	42.0	28.8	-	37.6	-	38.1
MM1.5-3B [96]	-	92.0	86.1	-	87.7	75.9	-	86.4	-	85.6
Robin-3B (Stage 1)	89.4	92.9	85.5	83.5	89.4	76.0	86.0	85.8	86.1	86.0
Robin-3B	91.1	93.9	87.7	86.1	91.4	79.8	88.3	88.4	**88.3**	**88.2**

Table 3. Comparison with SoTA models up to 13B parameters on Referring Expression Comprehension. The results are reported based on Recall@1 with IoU > 0.5.

Model	Open-Vocab. Segmentation			Region Classification			
	ADE [101]			LVIS [21]		PACO [65]	
	PQ	mAP	mIoU	SS	S-IOU	SS	S-IOU
LLaVA-1.5-7B [95]	-	-	-	48.9	19.8	42.2	14.6
Kosmos-2 [61]	6.5	4.3	5.4	38.9	8.7	32.1	4.8
Shikra-7B [11]	27.5	20.3	18.2	49.6	19.8	43.6	11.4
GPT4RoI-7B [97]	36.3	26.1	25.8	51.3	12.0	48.0	12.1
Ferret-7B [89]	39.5	29.9	31.8	63.8	36.6	58.7	26.0
Osprey-7B [92]	41.9	41.2	29.6	65.2	38.2	73.1	52.7
VisionLLMv2-Chat [80]	-	-	-	67.3	42.7	63.8	36.3
Robin-3B [Stage 1]	41.1	40.3	30.2	70.9	47.7	71.1	49.1
Robin-3B	**44.2**	**45.2**	**33.9**	**72.6**	**49.9**	**73.7**	**52.9**

Table 4. Results on Region Recognition. Following[92], we report panoptic segmentation (PQ), instance segmentation (mAP), and semantic segmentation (mIoU) on the ADE20K validation set. For region classification, we measure referring object classification on object-level LVIS and part-level PACO, reporting Semantic Similarity (SS) and Semantic Intersection over Union (S-IOU).

Method	# Relations	R@20	mR@20
Open-Ended Generation Models			
TextPSG [99]	50.0	4.8	-
ASM-V2-13B [78]	9.2	14.2	10.3
Robin-3B (Stage1)	5.6	18.2	10.9
Robin-3B	6.1	<u>20.6</u>	<u>13.2</u>
Closed-Set Detection Models			
IMP [81]	20.0	16.5	6.5
MOTIFS [93]	20.0	20.0	9.1
VC Tree [73]	20.0	20.6	9.7
GPSNet [45]	20.0	17.8	7.0
PSGFormer [84]	20.0	18.6	16.7
HiLO [102]	20.0	40.6	29.7
DSGG [22]	20.0	36.2	34.0

Table 5. Results on the Panoptic Scene Graph (PSG) Generation task. We report the recall (R@20) and mean recall (mR@20) of the predicted triplet relations. Underlined values denote the best results among open-ended generation models. Note that all closed-set models are single-task fine-tuned on the PSG dataset.

4.4. Scene graph generation

Lastly, we evaluate ROBIN on the task of scene graph detection using the Panoptic Scene Graph (PSG) dataset [84]. Here, the model must generate bounding boxes for object regions of interest and provide subject-predicate-object triplets for these detected regions. We prompt ROBIN to produce a complete scene graph, including bounding boxes, parse the output to extract object regions and predicates, and assign the text outputs to the closest object and predicate labels in the dataset via semantic similarity. We report Recall@20 (R@20) and mean Recall@20 (mR@20), which measure whether the ground-truth triplets appear in the top $K = 20$ predictions and match in bounding box (IoU > 0.5) and class labels.

We compare against MLMs that generate scene graphs as open-ended text [78, 99], as well as existing PSG detection-based models [45, 73, 81, 84, 93], which have been single-task fined-tuned on the PSG dataset to classify the relation triplets from a pre-defined closed set of class labels. Table 5 shows that our Robin-3B (Stage 1) model already outperforms previous open-ended generation models like ASM-V2-13B (18.2/10.9 vs. 14.2/10.3 on R@20/mR@20) After applying scene graph self-distillation in Stage 2, our Robin-3B model further improves to **20.6/13.2 on R@20/mR@20**. This demonstrates that scene graph self-distillation helps our model generate more accurate and diverse relation predictions. When comparing to closed-set detection models, our Robin-3B achieves an R@20 of 20.6, which is on par with early closed-set models like MOTIFS and VC Tree, both reporting an R@20 of around 20.0. Notably, these closed-set models are single-task fine-tuned on the PSG dataset and operate within a constrained set of relation classes, whereas our model operates in an open-ended setting without pre-defined set of relation classes. Overall, these results suggest the effectiveness of our approach in generating accurate scene graphs in an open-ended manner.

Qualitative Results Figure 6 shows a qualitative example of generated dense scene graphs in a complex scene. Notably, ROBIN correctly identifies different parts of objects with high accuracy, such as the interior dashboard [6], the door

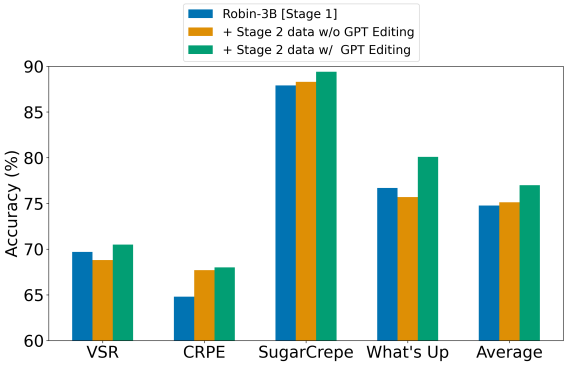

Figure 4. Effectiveness of SG-EDIT distillation.

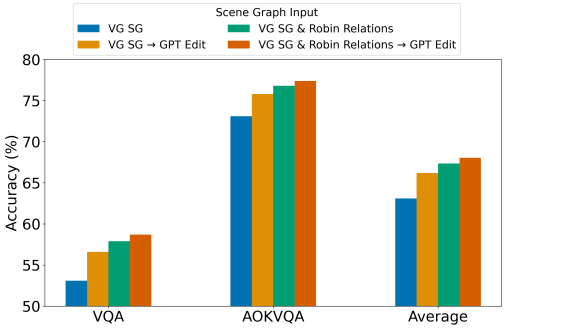

Figure 5. VQA performance of blind LLMs using different scene graphs as context: (1) original Visual Genome (VG) scene graphs, (2) GPT-edited VG scene graphs, (3) VG scene graphs appended with Robin-3B relations (Stage 1), and (4) the GPT-edited scene graphs from Stage 2.

[4], the front wheel [9], and the handle [20] as parts of the "silver convertible car with black top" [1]. It also includes spatial markers, such as the "left taillight" [17] and "right taillight" [18], as parts of the blue car [5]. Furthermore, it accurately associates the person [15] as working on the blue car [5], rather than other cars.

4.5. Ablation studies

In our ablation studies, we focus on showing the effectiveness of our proposed instruction tuning with dense scene graphs, and self-distillation with refinement from GPT4.

Training Dataset			Region Classification		Relationship Understanding	
			LVIS	PACO	VSR	CRPE
Grounding	Vis-Ins	SG	SS	SS	ZS-test	Relation
-	-	-	50.3	48.0	48.6	51.6
✓	-	-	49.3	47.1	48.1	54.0
✓	-	✓	55.3	53.5	48.9	60.4
-	✓	-	65.4	68.9	69.1	63.5
-	✓	✓	70.5	**72.5**	69.3	64.1
✓	✓	-	68.6	70.5	68.2	64.1
✓	✓	✓	**70.9**	71.1	**69.7**	**64.8**

Table 6. Ablation studies of instruction tuning with scene graph data. Different models trained with combinations of visual instruction (Vis-Ins), grounding (Grounding), and scene graph (SG) data, which are categorized in Stage 1 training (Section 3.2), are evaluated on region classification (SS: semantic similarity) and relationship understanding tasks. Note, the last row corresponds to the Robin-3B (Stage 1) model.

Role of scene graphs in instruction tuning To understand the benefits of incorporating scene graph data in visual understanding, we trained model variants using different combinations of the three datasets introduced in Stage 1 training (Sec 3.2), specifically comparing models trained with and without scene graph data. Table 6 shows the results. In region classification, we observe that adding scene graph (SG) data provides consistent gains across different training data configurations. Interestingly, the highest performance on the PACO dataset is achieved when training with only visual instruction and SG data (72.5 in SS). We suspect this is because the RefCOCO dataset is biased towards COCO objects rather than fine-grained, part-level objects. In contrast, our scene graph data encompass diverse relationships including part-level objects, providing enhanced reasoning capabilities for this task. Similarly, in relationship understanding tasks, we observe the same benefits of adding SG data with gains of 1.5% in VSR and 0.7% in CRPE.

Advantages of GPT4 editing and self-distillation Next, we explore the components introduced in our proposed SG-EDIT framework. We investigate the empirical gains of using GPT-4 edited scene graphs compared to training on the model's own generated scene graphs without refinement. Figure 4 compares the performance of ROBIN-3B trained with and without the GPT-4 edited scene graphs on relationship understanding benchmarks. We see a marginal average improvement of 0.5% over the Stage 1 model, when training the model additionally with its own scene graph generations (w/o GPT Editing). This gain is primarily due to improvements in the CRPE (+2.9%) and SugarCrepe (+0.4%) datasets, whereas performance decreases on the VSR (-0.9%) and What's Up (-1.0%) datasets. Meanwhile, we observe consistent gains for all benchmarks using edited scene graph (w/ edit) with average improvement of 2.2%.

Human-annotated vs. machine edited scene graphs Given the consistent gains from GPT-4 edited scene graphs, a natural question arises: can we simply refine human-annotated scene graphs with GPT-4 instead of relying on model-generated graphs? We hypothesize that a high-quality scene graph must encode sufficient visual details for a language model to accurately answer questions. To test this, we run an ablation study on VQAv2 [19] and AOKVQA [71] in a "blind LLM" setting, where the model has no direct image input but only a scene graph. As Figure 5 shows, editing scene graphs with GPT-4 (→ GPT Edit) consistently improves VQA performance compared to unedited versions. Notably, the model-generated scene graphs (VG SG & Robin Relations) outperform GPT-4 edits on human annotations alone (VG SG →GPT Edit), indicating that our model captures additional relationships missing from Visual Genome. Finally, applying GPT-4 edits on these model-generated graphs (VG SG & Robin Relations→ GPT Edit) leads to the highest accuracy, validating the proposed framework.

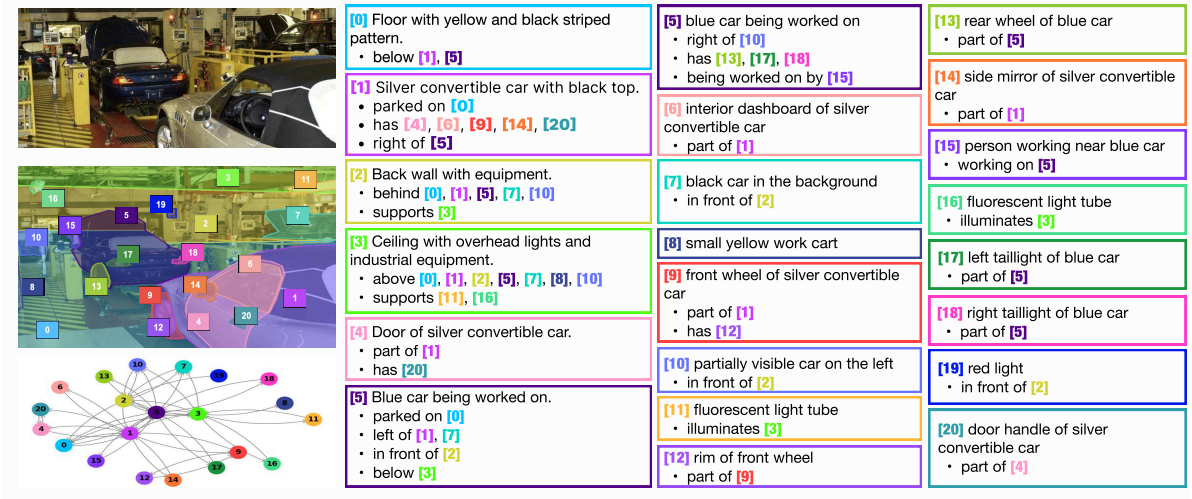

Figure 6. Qualitative example of dense scene graph generated by ROBIN-3B.

5. Related work

Relationship understanding and scene graphs Understanding relationships in visual scenes is a crucial challenge in computer vision, commonly addressed via scene graph generation (SGG) [81]. Compared to visual question answering [25, 46], scene graphs require a more comprehensive and structured understanding of the objects and their relationships within an image. Datasets such as Visual Relationship Detection [51] and Visual Genome [37] have stimulated research into SGG models, while Panoptic scene graph generation (PSG) [84] has extended bounding boxes based SGG to panoptic segmentation. Numerous end-to-end architectures rely on classification-based methods[41, 81, 91, 93, 99], and efforts exist to address object-relationship biases [74, 102] or to distill relationships from LLMs or MLMs [9, 35, 42]. However, these distillation-based approaches focus on improving the SGG task itself rather than enhancing the underlying MLMs.

Visual instruction tuning and MLMs The recent advancements in large language models [59] have resulted in the emergence of numerous LLM-based multimodal models [39, 47, 88, 103]. LLaVA is the first work that introduces visual instruction for LLMs, where the authors use language-only GPT-4 to generate a multi-modal instruction-following dataset [49]. Follow-up works have built other types of multi-modal instruction tuning data, including conversational-style QA [103], single-round QA based on academic datasets [15], detailed image descriptions [12].

Among the variants of MLMs, grounded MLMs [3, 10, 66, 76, 89, 100] are most relevant to our study, which are trained on region-based instruction tuning data to enable grounding capabilities of existing MLMs. A common representation of regions in such models is to refer the object regions as their bounding box coordinates in the text input-output [11, 95]. Alternatively, Osprey [92] is designed to understand a single or a handful of object segmentations at different levels of granularity. In this work, we extend Osprey to scene graph generation for a more comprehensive understanding of the entire image. ASM-v2 [78] is a grounded MLM designed for relation understanding and scene graph generation. Our approach relies on reasoning over relationships with diverse categories inferred by GPT4-V, and creates more complete and comprehensive scene graphs.

Data filtering of image-text datasets Radenovic et al. propose a rule-based system that filters out examples with low complexity [63], while LAION-400M introduces CLIP-filtering to remove image-text pairs with low similarity [64, 70]. Since then, various works have proposed improved filtering methods based on CLIP [16, 18]. A more recent work reports that finetuning MLMs such as LLaVa can yield better image-text data filters [77]. In contrast, we have developed a novel filtering pipeline that combines rule-based and model-based filtering for image-scene-graph data.

6. Conclusion

We present SVG and ROBIN to enhance the visual reasoning capabilities of MLMs. Our model outperforms the SoTA approaches in a suit of relationship reasoning tasks, grounded reasoning, and open-ended panoptic scene graph generation. Using the SG-EDIT distillation framework, ROBIN-3B model outperforms the state of the art 3B models in relationship understanding and REC tasks, and even 13B models trained with a similar amount of data. One limitation is the lack of evaluation of scene graph generations beyond existing annotations on COCO images but for images in the wild, which we leave it as future work. Another extension is to include reasoning over the shapes and structures of 3D objects, enhancing consistency in video understanding, and facilitating image generation with controllable layouts.

References

[1] Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, et al. Phi-3 technical report: A highly capable language model locally on your phone. arXiv preprint arXiv:2404.14219, 2024. 4, 5, 6

[2] Manoj Acharya, Kushal Kafle, and Christopher Kanan. Tallyqa: Answering complex counting questions. In AAAI Conference on Artificial Intelligence, 2018. 4, 5

[3] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. arXiv preprint arXiv:2308.12966, 2023. 5, 6, 8, 7

[4] Irving Biederman. Recognition-by-components: a theory of human image understanding. Psychological review, 94(2):115, 1987. 1

[5] Irving Biederman, Robert J Mezzanotte, and Jan C Rabinowitz. Scene perception: Detecting and judging objects undergoing relational violations. Cognitive psychology, 14(2):143–177, 1982. 1

[6] Léon Bottou. From machine learning to machine reasoning. Machine learning, 94(2):133–149, 2014. 1

[7] Zheng Cai, Maosong Cao, Haojiong Chen, Kai Chen, Keyu Chen, Xin Chen, Xun Chen, Zehui Chen, Zhi Chen, Pei Chu, et al. Internlm2 technical report. arXiv preprint arXiv:2403.17297, 2024. 4, 5, 6

[8] Boyuan Chen, Zhuo Xu, Sean Kirmani, Brain Ichter, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 14455–14465, 2024. 1

[9] Guikun Chen, Jin Li, and Wenguan Wang. Scene graph generation with role-playing large language models. In NeurIPS, 2024. 8

[10] Jun Chen, Deyao Zhu, Xiaoqian Shen, Xiang Li, Zechun Liu, Pengchuan Zhang, Raghuraman Krishnamoorthi, Vikas Chandra, Yunyang Xiong, and Mohamed Elhoseiny. Minigpt-v2: large language model as a unified interface for vision-language multi-task learning. arXiv preprint arXiv:2310.09478, 2023. 5, 6, 8

[11] Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. Shikra: Unleashing multimodal llm's referential dialogue magic. arXiv preprint arXiv:2306.15195, 2023. 4, 5, 6, 8

[12] 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. 5, 8

[13] 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

[14] Noam Chomsky and Morris Halle. Some controversial questions in phonological theory. Journal of linguistics, 1(2):97–138, 1965. 1

[15] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. Advances in Neural Information Processing Systems, 36, 2024. 1, 5, 8

[16] Alex Fang, Albin Madappally Jose, Amit Jain, Ludwig Schmidt, Alexander Toshev, and Vaishaal Shankar. Data filtering networks. arXiv preprint arXiv:2309.17425, 2023. 3, 8

[17] Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A Smith, Wei-Chiu Ma, and Ranjay Krishna. Blink: Multimodal large language models can see but not perceive. arXiv preprint arXiv:2404.12390, 2024. 2

[18] Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, et al. Datacomp: In search of the next generation of multimodal datasets. Advances in Neural Information Processing Systems, 36, 2024. 8

[19] 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. 7, 4, 5

[20] Madeleine Grunde-McLaughlin, Ranjay Krishna, and Maneesh Agrawala. Agqa: A benchmark for compositional spatio-temporal reasoning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021. 1

[21] Agrim Gupta, Piotr Dollar, and Ross Girshick. Lvis: A dataset for large vocabulary instance segmentation. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 5356–5364, 2019. 3, 5, 6

[22] Zeeshan Hayder and Xuming He. Dsgg: Dense relation transformer for an end-to-end scene graph generation. In Proceedings of the IEEE/CVF Conference

on Computer Vision and Pattern Recognition, pages 28317–28326, 2024. 6

[23] Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration, 2020. 4

[24] Cheng-Yu Hsieh, Jieyu Zhang, Zixian Ma, Aniruddha Kembhavi, and Ranjay Krishna. Sugarcrepe: Fixing hackable benchmarks for vision-language compositionality. Advances in Neural Information Processing Systems, 36, 2024. 4, 5, 7

[25] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 6700–6709, 2019. 3, 4, 5, 8

[26] Dieuwke Hupkes, Verna Dankers, Mathijs Mul, and Elia Bruni. Compositionality decomposed: How do neural networks generalise? Journal of Artificial Intelligence Research, 67:757–795, 2020. 1

[27] Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. Openclip, 2021. If you use this software, please cite it as below. 4, 3

[28] Theo MV Janssen and Barbara H Partee. Compositionality. In Handbook of logic and language, pages 417–473. Elsevier, 1997. 1

[29] Jingwei Ji, Ranjay Krishna, Li Fei-Fei, and Juan Carlos Niebles. Action genome: Actions as compositions of spatio-temporal scene graphs. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 10236–10247, 2020. 1

[30] Kushal Kafle, Brian Price, Scott Cohen, and Christopher Kanan. Dvqa: Understanding data visualizations via question answering. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 5648–5656, 2018. 4

[31] Amita Kamath, Jack Hessel, and Kai-Wei Chang. What's "up" with vision-language models? investigating their struggle with spatial reasoning. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. 4

[32] Amita Kamath, Jack Hessel, and Kai-Wei Chang. What's" up" with vision-language models? investigating their struggle with spatial reasoning. arXiv preprint arXiv:2310.19785, 2023. 5

[33] Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A diagram is worth a dozen images. In Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14, pages 235–251. Springer, 2016. 4, 5

[34] Aniruddha Kembhavi, Minjoon Seo, Dustin Schwenk, Jonghyun Choi, Ali Farhadi, and Hannaneh Hajishirzi. Are you smarter than a sixth grader? textbook question answering for multimodal machine comprehension. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 5376–5384, 2017. 4, 5

[35] Kibum Kim, Kanghoon Yoon, Jaeyeong Jeon, Yeonjun In, Jinyoung Moon, Donghyun Kim, and Chanyoung Park. Llm4sgg: Large language models for weakly supervised scene graph generation. 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 28306–28316, 2023. 8

[36] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 4015–4026, 2023. 3, 7

[37] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. International journal of computer vision, 123:32–73, 2017. 1, 2, 3, 4, 8, 5, 6, 14

[38] Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. Seed-bench: Benchmarking multimodal llms with generative comprehension, 2023. 4, 5

[39] Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Fanyi Pu, Jingkang Yang, Chunyuan Li, and Ziwei Liu. Mimic-it: Multi-modal in-context instruction tuning. arXiv preprint arXiv:2306.05425, 2023. 8

[40] Feng Li, Hao Zhang, Peize Sun, Xueyan Zou, Shilong Liu, Jianwei Yang, Chunyuan Li, Lei Zhang, and Jianfeng Gao. Semantic-sam: Segment and recognize anything at any granularity. arXiv preprint arXiv:2307.04767, 2023. 3, 7

[41] Rongjie Li, Songyang Zhang, and Xuming He. Sgtr: End-to-end scene graph generation with transformer, 2022. 8

[42] Rongjie Li, Songyang Zhang, Dahua Lin, Kai Chen, and Xuming He. From pixels to graphs: Open-vocabulary scene graph generation with vision-language models. 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 28076–28086, 2024. 8

[43] Ji Lin, Hongxu Yin, Wei Ping, Pavlo Molchanov, Mohammad Shoeybi, and Song Han. Vila: On pre-

training for visual language models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 26689–26699, 2024. 4, 5

[44] 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. 3

[45] 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. 6

[46] Fangyu Liu, Guy Emerson, and Nigel Collier. Visual spatial reasoning. Transactions of the Association for Computational Linguistics, 11:635–651, 2023. 4, 5, 8

[47] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. arXiv preprint arXiv:2310.03744, 2023. 8, 4

[48] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, 2024. 5

[49] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. Advances in neural information processing systems, 36, 2024. 8, 7

[50] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? arXiv preprint arXiv:2307.06281, 2023. 4, 5

[51] 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, 8

[52] Chuofan Ma, Yi Jiang, Jiannan Wu, Zehuan Yuan, and Xiaojuan Qi. Groma: Localized visual tokenization for grounding multimodal large language models. In European Conference on Computer Vision, pages 417–435. Springer, 2025. 4, 6, 3

[53] Zixian Ma, Jerry Hong, Mustafa Omer Gul, Mona Gandhi, Irena Gao, and Ranjay Krishna. Crepe: Can vision-language foundation models reason compositionally? In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 10910–10921, 2023. 4, 7

[54] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 11–20, 2016. 5, 6

[55] Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge. In Proceedings of the IEEE/cvf conference on computer vision and pattern recognition, pages 3195–3204, 2019. 4, 5

[56] Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. arXiv preprint arXiv:2203.10244, 2022. 4

[57] Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. Docvqa: A dataset for vqa on document images. In Proceedings of the IEEE/CVF winter conference on applications of computer vision, pages 2200–2209, 2021. 4

[58] Brandon McKinzie, Zhe Gan, Jean-Philippe Fauconnier, Sam Dodge, Bowen Zhang, Philipp Dufter, Dhruti Shah, Xianzhi Du, Futang Peng, Floris Weers, et al. Mm1: Methods, analysis & insights from multimodal llm pre-training. arXiv preprint arXiv:2403.09611, 2024. 5

[59] et al. OpenAI. Gpt-4 technical report, 2024. 8

[60] Jae Sung Park, Jack Hessel, Khyathi Raghavi Chandu, Paul Pu Liang, Ximing Lu, Peter West, Youngjae Yu, Qiuyuan Huang, Jianfeng Gao, Ali Farhadi, and Yejin Choi. Localized symbolic knowledge distillation for visual commonsense models. ArXiv, abs/2312.04837, 2023. 3

[61] 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. 6

[62] Bryan A. Plummer, Liwei Wang, Chris M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models, 2016. 4

[63] Filip Radenovic, Abhimanyu Dubey, Abhishek Kadian, Todor Mihaylov, Simon Vandenhende, Yash Patel, Yi Wen, Vignesh Ramanathan, and Dhruv Mahajan. Filtering, distillation, and hard negatives for vision-language pre-training. In Proceedings of

the IEEE/CVF conference on computer vision and pattern recognition, pages 6967–6977, 2023. 8

[64] Alec Radford, Jong Wook Kim, Chris Hallacy, A. Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In ICML, 2021. 8

[65] Vignesh Ramanathan, Anmol Kalia, Vladan Petrovic, Yi Wen, Baixue Zheng, Baishan Guo, Rui Wang, Aaron Marquez, Rama Kovvuri, Abhishek Kadian, et al. Paco: Parts and attributes of common objects. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 7141–7151, 2023. 5, 6

[66] Hanoona Rasheed, Muhammad Maaz, Sahal Shaji, Abdelrahman Shaker, Salman Khan, Hisham Cholakkal, Rao M Anwer, Erix Xing, Ming-Hsuan Yang, and Fahad S Khan. Glamm: Pixel grounding large multimodal model. arXiv preprint arXiv:2311.03356, 2023. 8

[67] Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pages 3505–3506, 2020. 3

[68] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. arXiv preprint arXiv:1908.10084, 2019. 5, 6

[69] Tanik Saikh, Tirthankar Ghosal, Amish Mittal, Asif Ekbal, and Pushpak Bhattacharyya. Scienceqa: A novel resource for question answering on scholarly articles. International Journal on Digital Libraries, 23 (3):289–301, 2022. 4, 5

[70] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. arXiv preprint arXiv:2111.02114, 2021. 8, 7

[71] Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. A-okvqa: A benchmark for visual question answering using world knowledge. arXiv, 2022. 7, 4, 5

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

[73] 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. 6

[74] Kaihua Tang, Yulei Niu, Jianqiang Huang, Jiaxin Shi, and Hanwang Zhang. Unbiased scene graph generation from biased training. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 3716–3725, 2020. 8

[75] Jiayu Wang, Yifei Ming, Zhenmei Shi, Vibhav Vineet, Xin Wang, Yixuan Li, and Neel Joshi. Is a picture worth a thousand words? delving into spatial reasoning for vision language models. arXiv preprint arXiv:2406.14852, 2024. 1

[76] Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, et al. Cogvlm: Visual expert for pretrained language models. arXiv preprint arXiv:2311.03079, 2023. 8, 7

[77] Weizhi Wang, Khalil Mrini, Linjie Yang, Sateesh Kumar, Yu Tian, Xifeng Yan, and Heng Wang. Finetuned multimodal language models are high-quality image-text data filters. arXiv preprint arXiv:2403.02677, 2024. 8

[78] Weiyun Wang, Yiming Ren, Haowen Luo, Tiantong Li, Chenxiang Yan, Zhe Chen, Wenhai Wang, Qingyun Li, Lewei Lu, Xizhou Zhu, Yu Qiao, and Jifeng Dai. The all-seeing project v2: Towards general relation comprehension of the open world, 2024. 2, 4, 5, 6, 8

[79] Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Anjana Arunkumar, Arjun Ashok, Arut Selvan Dhanasekaran, Atharva Naik, David Stap, et al. Super-naturalinstructions: Generalization via declarative instructions on 1600+ nlp tasks. arXiv preprint arXiv:2204.07705, 2022. 2

[80] Jiannan Wu, Muyan Zhong, Sen Xing, Zeqiang Lai, Zhaoyang Liu, Wenhai Wang, Zhe Chen, Xizhou Zhu, Lewei Lu, Tong Lu, et al. Visionllm v2: An end-to-end generalist multimodal large language model for hundreds of vision-language tasks. arXiv preprint arXiv:2406.08394, 2024. 5, 6

[81] 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. 6, 8

[82] Le Xue, Manli Shu, Anas Awadalla, Jun Wang, An Yan, Senthil Purushwalkam, Honglu Zhou, Viraj Prabhu, Yutong Dai, Michael S Ryoo, et al. xgen-mm

(blip-3): A family of open large multimodal models. arXiv preprint arXiv:2408.08872, 2024. 4, 5

[83] An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zhihao Fan. Qwen2 technical report. arXiv preprint arXiv:2407.10671, 2024. 4

[84] 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. 3, 4, 6, 8, 5

[85] Jianwei Yang, Hao Zhang, Feng Li, Xueyan Zou, Chunyuan Li, and Jianfeng Gao. Set-of-mark prompting unleashes extraordinary visual grounding in gpt-4v. arXiv preprint arXiv:2310.11441, 2023. 2

[86] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In CVPR, 2024. 3

[87] Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, et al. Minicpm-v: A gpt-4v level mllm on your phone. arXiv preprint arXiv:2408.01800, 2024. 4, 5

[88] Qinghao Ye, Haiyang Xu, Jiabo Ye, Ming Yan, Haowei Liu, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration. arXiv preprint arXiv:2311.04257, 2023. 8

[89] Haoxuan You, Haotian Zhang, Zhe Gan, Xianzhi Du, Bowen Zhang, Zirui Wang, Liangliang Cao, Shih-Fu Chang, and Yinfei Yang. Ferret: Refer and ground anything anywhere at any granularity. arXiv preprint arXiv:2310.07704, 2023. 6, 8

[90] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14, pages 69–85. Springer, 2016. 3, 5, 6, 4

[91] Hangjie Yuan, Jianwen Jiang, Samuel Albanie, Tao Feng, Ziyuan Huang, Dong Ni, and Mingqian Tang. Rlip: Relational language-image pre-training for human-object interaction detection, 2022. 8

[92] Yuqian Yuan, Wentong Li, Jian Liu, Dongqi Tang, Xinjie Luo, Chi Qin, Lei Zhang, and Jianke Zhu. Osprey: Pixel understanding with visual instruction tuning, 2024. 4, 5, 6, 8, 3

[93] 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. 6, 8

[94] Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. From recognition to cognition: Visual commonsense reasoning. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 6720–6731, 2019. 4

[95] Hao Zhang, Hongyang Li, Feng Li, Tianhe Ren, Xueyan Zou, Shilong Liu, Shijia Huang, Jianfeng Gao, Lei Zhang, Chunyuan Li, et al. Llava-grounding: Grounded visual chat with large multimodal models. arXiv preprint arXiv:2312.02949, 2023. 4, 5, 6, 8

[96] Haotian Zhang, Mingfei Gao, Zhe Gan, Philipp Dufter, Nina Wenzel, Forrest Huang, Dhruti Shah, Xianzhi Du, Bowen Zhang, Yanghao Li, et al. Mm1. 5: Methods, analysis & insights from multimodal llm fine-tuning. arXiv preprint arXiv:2409.20566, 2024. 4, 5, 6

[97] Shilong Zhang, Peize Sun, Shoufa Chen, Min Xiao, Wenqi Shao, Wenwei Zhang, Yu Liu, Kai Chen, and Ping Luo. Gpt4roi: Instruction tuning large language model on region-of-interest. arXiv preprint arXiv:2307.03601, 2023. 6, 3

[98] Yanzhe Zhang, Ruiyi Zhang, Jiuxiang Gu, Yufan Zhou, Nedim Lipka, Diyi Yang, and Tong Sun. Llavar: Enhanced visual instruction tuning for text-rich image understanding. arXiv preprint arXiv:2306.17107, 2023. 4

[99] 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. 6, 8

[100] Yang Zhao, Zhijie Lin, Daquan Zhou, Zilong Huang, Jiashi Feng, and Bingyi Kang. Bubogpt: Enabling visual grounding in multi-modal llms. arXiv preprint arXiv:2307.08581, 2023. 8

[101] Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Semantic understanding of scenes through the ade20k dataset. International Journal of Computer Vision, 127:302–321, 2019. 3, 5, 6, 1

[102] Zijian Zhou, Miaojing Shi, and Holger Caesar. Hilo: Exploiting high low frequency relations for unbiased panoptic scene graph generation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 21637–21648, 2023. 6, 8

[103] 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. 8