GSR-Bench: A Benchmark for Grounded Spatial Reasoning Evaluation via Multimodal LLMs

Anonymous ACL submission

Abstract

The ability to understand and reason about spatial relationships between objects in images is an important component of visual reasoning. This skill rests on the ability to recognize and localize objects of interest and determine their spatial relation. Early vision and language models (VLMs) have been shown to struggle to recognize spatial relations. We extend the previously released What'sUp dataset (Kamath et al., 2023) and propose a novel comprehensive evaluation for spatial relationship understanding that highlights the strengths and weaknesses of 27 different models. In addition to the VLMs evaluated in What'sUp, our extensive evaluation encompasses 3 classes of Multimodal LLMs (MLLMs) that vary in their parameter sizes (ranging from 7B to 110B), training/instruction-tuning methods, and visual resolution to benchmark their performances and scrutinize the scaling laws in this task.

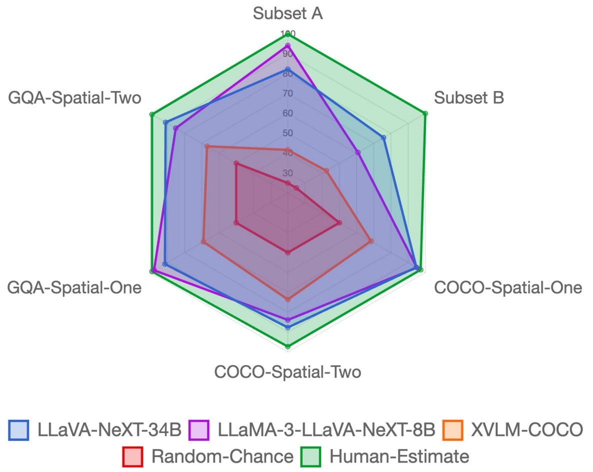


Figure 1: LLaMA-3-LLaVA-NeXT-8B achieves the overall accuracy of 86.1%, compared to 60.4% by XVLM-COCO, in What'sUp benchmark, reaching the best trade-off between accuracy and parameters size, since it performs only 1.1% lower than LLaVA-NeXT-34B, which has $\times 4.25$ number of parameters.

1 Introduction

Earlier efforts for benchmarking vision and language models (VLMs) were developed for cross-modal and/or dual-encoder, end-to-end models, like LXMERT (Tan and Bansal, 2019), CLIP (Radford et al., 2021), BLIP (Li et al., 2022), with the focus on downstream tasks performances such as VQA (Antol et al., 2015), GQA (Hudson and Manning, 2019), referring expressions (Kazemzadeh et al., 2014), image-text matching or image/text retrieval. While spatial relations are often part of VQA datasets, the evaluation of spatial reasoning is often conflated with grounding referring expressions or objects and their attributes[1]. To isolate these issues, authors in (Kamath et al., 2023) introduced a new benchmark that focuses on spatial relationship understanding only. Using image-text matching evaluation methodology, they showed that contrastive models such as CLIP, BLIP, and their follow-up variants struggle to understand spatial relations with the best accuracy around 61%.

Recent advances in generative large multi-modal models have shown remarkable visual knowledge and reasoning capabilities. We revisit the spatial relationship understanding in the context of MLLMs and extend the existing What'sUp benchmark (Kamath et al., 2023) to include bounding box annotations and depth information. Compositional spatial relationship understanding requires successful recognition of objects and determining their locations. Furthermore, the knowledge of scene depth helps to disambiguate certain relationships (e.g., "*in front of*" or "*behind*"). The availability of this information can support a grounded understanding of spatial relations and will contribute to the fine-grained evaluation of large generative MLLMs, which lag behind their earlier counterparts. A few exceptions are multi-task multi-modal

[1] VQA example question may be: "Is there a woman to the left of the person that is wearing a wetsuit?"

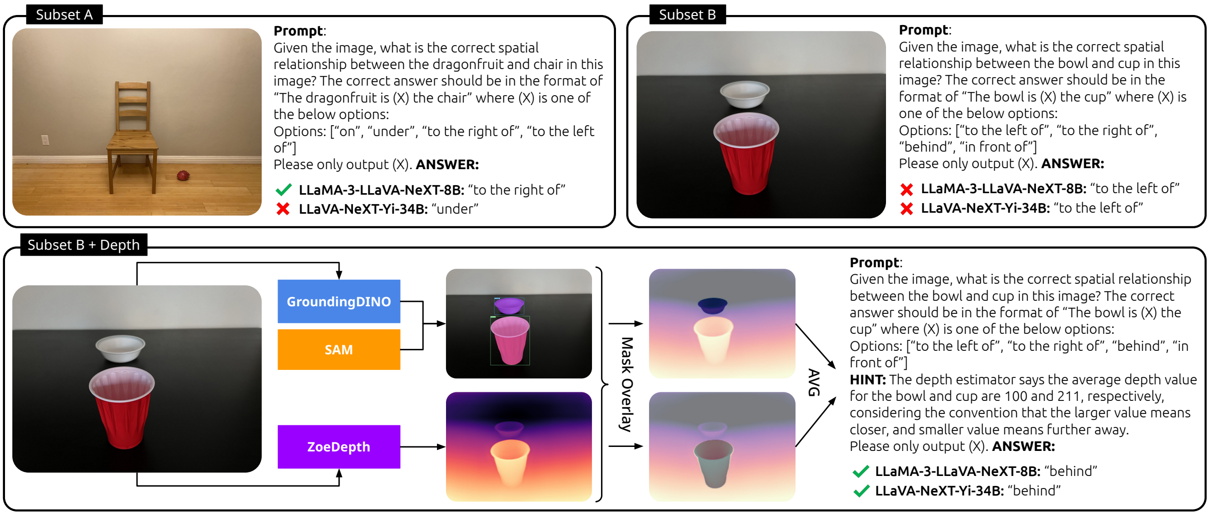


Figure 2: Our pipeline overview for spatial relationship understanding prompting, shown in the top two figures, and our depth-augmented prompting, shown in the bottom figure.

benchmarks like MMBench (Liu et al., 2023d) and its related benchmarks that focus on evaluating several MLLMs for both visual recognition tasks and description generation. Given the simple structure of spatial clauses, we can study separately the ability of the model to ground the subject and object in the clause, and the effect and means of incorporating the depth information. The contributions of this work can be summarized as follows:

- Extended What'sUp spatial relationship dataset with depth, masks, and bounding box annotations.
- Design of different prompting strategies through structured prompting for the evaluation of grounding and spatial reasoning.
- Comprehensive evaluation and comparison of 18 VLMs and 9 MLLMs, with various sizes, resolutions, pre-training/instruction-tuning, and prompting strategies.

2 GSR Benchmark

We extend carefully curated What'sUp dataset (Kamath et al., 2023) that is comprised of Subset A containing pairs of objects in unambiguous spatial relations, being "*on*", "*under*", "*left of*" or "*right of*" a table, chair, or armchair, and Subset B containing an object "*in front of*", "*behind*", "*left*" or "*right*" of another object on a tabletop, and subsets of COCO-Spatial and GQA-Spatial with either one or two objects occurring, accompanied by spatial clauses like "*on top of*", "*on the bottom of*", "*right of*", or "*left of*". To study the grounding in this context, we annotate the dataset with bounding box coordinates and segmentation masks for all the objects mentioned in the captions and the depth maps for the images. We leverage GroundingDINO (Liu et al., 2023c) as an open-vocabulary object detector, Segment Anything (SAM) (Kirillov et al., 2023) for the object mask segmentation, and ZoeDepth (Bhat et al., 2023) for monocular depth estimation. In the next section, we explain in detail how these additional annotations enable a more rigorous and grounded evaluation of spatial reasoning and its components[2].

3 GSR-Bench Experiments

Grounded spatial reasoning evaluation is typically done using image-text matching, binary VQA, or multiple-choice VQA. Further evaluations include subject and/or object grounding and localization; and exploring the effect of using depth information. In addition to 18 VLMs that have been evaluated in (Kamath et al., 2023), we focus on the probing of open-source generative MLLMs like LLaVA and InternVL[3] using structured generation methodologies of Multiple choice (MC) and Template-based generation (TG). In MC prompting, captions for each image are represented as A, B, C, and D options for Subset A and Subset B, while A and B options for COCO-Spatial and GQA-Spatial Subsets. Then, the model is prompted to choose the correct letter as the final answer. In TG prompting,

[2] All the code and data will be publicly available.
[3] InternVL is the leading model in MMBench.

Model	Num Params	Subset A Sub-Obj	Subset B Sub-Obj	COCO-spatial One-Obj	COCO-spatial Two-Obj	GQA-spatial One-Obj	GQA-spatial Two-Obj	Total Average
CLIP ViT-B/32 (Radford et al., 2021)	151M	30.3	31.6	43.7	51.1	46.5	47.4	41.8
CLIP ViT-L/14	428M	26.5	25.7	49.2	49.8	46.1	48.5	41.0
NegCLIP (Yuksekgonul et al., 2022)	–	32.5	36.3	47.4	46.4	45.3	46.7	42.4
RoBERTaCLIP (Kamath et al., 2023)	–	25.2	25.0	46.3	53.6	50.8	48.8	41.6
CoCa (Yu et al., 2022)	2.1B	29.4	29.4	48.1	45.2	45.0	49.1	41.0
XVLM 4M (Zeng et al., 2021)	216M	40.0	23.0	58.4	65.0	62.8	54.6	50.6
XVLM 16M	216M	50.7	33.1	65.4	64.5	63.2	53.3	55.0
BLIP 14M (Li et al., 2022)	583M	38.8	38.2	54.2	53.9	49.1	50.5	47.5
BLIP 129M	583M	30.3	30.4	44.8	53.9	50.5	47.4	42.9
BLIP2-ITM (Li et al., 2023)	188M	44.9	30.4	48.3	57.7	46.0	53.6	46.8
BLIP2-ITC	188M	35.9	22.1	55.6	51.8	52.6	49.5	44.6
FLAVA (Singh et al., 2022)	–	33.7	27.2	50.3	55.0	52.2	51.2	44.9
CoCa-Caption	2.1B	25.5	22.8	45.9	51.4	48.5	50.5	40.8
XVLM-Flickr30K	216M	45.1	*43.4*	63.1	67.3	64.7	58.1	56.9
XVLM-COCO	216M	41.7	42.4	*68.4*	*73.6*	*69.1*	*67.0*	*60.4*
BLIP-Flickr30K	583M	29.6	38.0	50.0	58.4	50.3	47.4	45.6
BLIP-COCO	583M	35.7	29.9	46.4	56.4	50.3	52.6	45.2
BLIP-VQA	583M	*57.8*	37.7	63.6	60.5	63.8	52.9	56.0
LLaVA-1.5-Vicuna	7B	25.0	31.9	90.4	66.6	91.2	62.9	61.3
LLaVA-1.5-Vicuna	13B	58.5	28.2	92.5	78.9	93.1	82.8	72.3
LLaVA-NeXT-Mistral	7B	37.4	22.0	81.1	60.4	89.4	57.0	57.9
LLaVA-NeXT-Vicuna	7B	38.6	26.2	95.5	71.8	97.6	79.0	68.1
LLaVA-NeXT-Vicuna	13B	75.0	20.1	**95.6**	78.6	97.6	84.9	75.3
LLaMA-3-LLaVA-NeXT	8B	**94.2**	60.8	95.1	83.9	**97.8**	85.2	86.1
LLaVA-NeXT-Yi	34B	82.3	**75.7**	94.8	**87.7**	91.5	91.1	**87.2**
LLaVA-NeXT-Qwen1.5	110B	93.9	54.2	90.6	84.1	96.2	**94.2**	85.4
Intern-VL-Chat-1.5	26B	92.2	61.8	95.1	82.3	**97.8**	82.8	85.3
Random Chance	–	25.0	25.0	50.0	50.0	50.0	50.0	41.7

Table 1: Template-based generation (TG) results using CircularEval. The first two sections come from What'sUp (Kamath et al., 2023) results. The rest shows our LLaVA 1.5, 1.6, and InternVL-1.5 prompting results. Our best-performing is shown in **bold**, 2nd-best with underline, and What'sUp best-performing with *italic underline*.

as shown in Figure 3, we append the correct format of the entire caption to the prompt, in which the spatial clause acts as the placeholder for the correct spatial relation option. In this way, we are able to leverage LLMs' open-ended generation capability, handle the models' verbosity by enforcing the correct answer structure, and overcome the biases observed in MC prompting simultaneously (See Figure 2).

Sample Prompt

Given the image, what is the correct spatial relationship between the **subject** and **object** in this image? The correct answer should be in the format of "The **subject** is (X) the **object**.", where (X) is one of the below options:
Options: ["on", "under", "to the right of", "to the left of"]
Please only output (X), without any other output.
ANSWER:

Figure 3: TG sample prompt structure.

We ran each prompt with 4 different permutations so as to vary the position of the answer among the choices in MC and the list of options in TG prompting. An instance is considered correct if all four options are predicted correctly, known as *CircularEval*, introduced in MMBench (Liu et al., 2023d). As opposed to the CircularEval, there exists VanillaEval, which only asks the model to choose the correct answer from a list of options once and has been shown to be prone to bias in recent studies. We first ran our experiments using MC prompting and observed a significant degree of bias among the models when the position of the answer varied among the choices of A, B, C, or D. This bias and sensitivity turned out to be even more detrimental in smaller models, while larger models like LLaVA-NeXT-Yi-34B and LLaMA-3-LLaVA-NeXT-8B showed significantly higher robustness (See Figure 4 in the Appendix for details). This phenomenon also corroborates the findings of multiple recent studies in LLMs (Zheng et al., 2023; Pezeshkpour and Hruschka, 2023; Wang et al., 2023; Xue et al., 2024; Wang et al., 2024). According to this observation, we opted for TG prompting, accompanied by the CircularEval methodology, inspired by Gemini 1.5 Pro (Reid et al., 2024). See Table 1 for the TG prompting results, where rows in section 1 and 2 come from the What'sUp benchmark (Kamath et al., 2023), section 3 refers to LLaVA-1.5 models (Liu et al., 2023b), section 4 to the LLaVA-NeXT models (Li et al., 2024; Liu et al., 2024), and section 5 to the

Model	Subset A		Subset B		COCO-spatial			GQA-spatial			Avg
	Sub	Obj	Sub	Obj	One-Obj	Sub	Obj	One-Obj	Sub	Obj	G-Score
LLaVA-1.5-Vicuna-7B	9.7	79.4	51.5	25.7	47.4	49.8	48.0	31.9	55.0	47.8	44.62
LLaVA-1.5-Vicuna-13B	13.8	86.1	77.4	32.3	60.9	61.8	61.0	42.1	72.2	59.8	56.74
LLaVA-NeXT-Vicuna-7B	14.1	99.0	95.8	66.7	81.9	84.5	77.7	45.5	60.1	56.0	68.13
LLaVA-NeXT-Mistral-7B	13.1	82.3	93.9	60.0	<u>87.1</u>	86.8	<u>85.7</u>	69.2	85.9	81.8	74.58
LLaVA-NeXT-Vicuna-13B	15.3	84.0	95.3	67.6	<u>87.1</u>	**90.2**	83.9	69.6	85.9	80.4	75.93
LLaMA-3-LLaVA-NeXT-8B	19.2	<u>99.3</u>	96.6	73.8	85.7	87.5	83.2	69.0	84.5	80.4	77.92
LLaVA-NeXT-Yi-34B	<u>21.1</u>	**100.0**	<u>97.8</u>	<u>78.9</u>	83.7	85.7	81.4	<u>70.0</u>	**88.0**	<u>83.5</u>	<u>79.01</u>
LLaVA-NeXT-Qwen1.5-110B	**29.4**	98.5	**98.8**	**80.1**	**88.7**	<u>88.2</u>	**86.4**	**74.9**	<u>86.9</u>	**84.2**	**81.61**
GroundingDINO [avg(ρ)]	58.8	92.0	78.1	70.1	62.3	62.8	59.3	59.4	70.4	65.2	67.84
GroundingDINO [$\Sigma(\rho \geq 0.5)/t$]	68.9	100.0	90.0	88.7	71.0	73.6	66.4	59.1	76.3	71.1	76.51

Table 2: Grounding/Localization results. Avg G-Score refers to the mean accuracy of IoU ≥ 0.5. The bottom two rows refer to the GroundingDINO mean confidence scores (ρ), and mean accuracy of $\rho \geq 0.5$, respectively.

InternVL-1.5 results (Chen et al., 2024).

Grounding/Localization Evaluation. This experiment aims to measure the MLLMs grounding ability of the objects mentioned in the captions. Recent studies like (Rajabi and Kosecka, 2023) on Visual Spatial Reasoning (VSR) benchmark (Liu et al., 2023a) has demonstrated that there exist multiple cases where the VLM correctly predicts the binary ITM label of 1 using the holistic representations of the image and caption, while the model fails to localize the subject and object correctly. Our experiments aim to quantify these type of behaviors in MLLMs. We prompt MLLMs to extract the normalized bounding box coordinates for the caption's objects as "`Give me the bounding box coordinates for the {object}`" and compute the IoU between the model's output and the GroundingDINO output for each object, assigning the binary accuracy of 1 if IoU ≥ 0.5, otherwise 0. See Table 2 for the results.

Model	w/o depth	with depth
InternVL-Chat-1.5-26B	26.5	40.7
LLaMA-3-LLaVA-NeXT-8B	53.4	60.3
LLaVA-NeXT-Yi-34B	64.7	81.9

Table 3: DAP results for *behind* & *in front of* cases.

Depth-Augmented Prompting (DAP). The experiments in Table 1 revealed that Subset B is the lowest-performing, with many instances requiring reasoning about "*behind*" and "*in front of*" spatial clauses. We propose to incorporate the depth values of `subject` and `object` into the prompt, as a hint to the model, utilizing our augmented benchmark annotations, depicted in Figure 2. We show that this minimal change improves the accuracy of top-3 performing models in these instances of Subset B, reported by CircularEval in Table 3.

4 Discussion

According to Table 1 and 2, there is a positive correlation, even stronger in grounding, between **scaling the LLM size** & **visual resolution**, and the **overall accuracy** in both tasks. Conversely, there exist multiple exceptions, which are inevitable to concretely justify due to various intervening factors, such as (1) differences in training/fine-tuning & architectures and (2) release date and further instruction-tuning of the LLMs, like LLaMA-3-8B, which has the most-recent knowledge cut-off.

Grounding small objects, which refers to the Sub column in Subset A, seems challenging for all, and worst in smaller models, according to Table 2. We also observed a plateau in Table 1, especially in Qwen-1.5-110B, which is the largest ever released open-source MLLM at the moment. This could be a sign of saturation where the reasoning capability flattens out, although scaling still improves grounding, shown in Table 2.

5 Conclusions

In this work, we introduce a new benchmark for grounded spatial reasoning by enriching the What'sUp dataset with additional supervision for a more fine-grained assessment of MLLM's spatial understanding. We also propose a new compositional evaluation methodology for (1) a stricter assessment of spatial relationship understanding through CircularEval, and (2) measuring the model's grounding capability using the labels we generate through our cost-effective auto-annotation pipeline. Our evaluations reveal the superiority of LLaVA MLLMs over the best-performing VLMs evaluated in What'sUp, like XVLM, by a significant margin of $\sim +26.8\%$. Future works may investigate the remaining gap between the top open-source MLLMs and human-level accuracy.

Limitations

Small-scale Dataset: Our split sizes remain the same as the What'sUp dataset in which Subset A has 412, Subset B has 408, COCO-Spatial-One has 2247, COCO-Spatial-Two has 440, GQA-Spatial-One has 1160, and GQA-Spatial-Two has 291 instances. Although this benchmark includes 4,958 image instances in total, each instance covering one or two objects, with various domain shifts in each 6 split, it is smaller than already existing benchmarks related to spatial reasoning, like Visual Genome (Krishna et al., 2017), GQA (Hudson and Manning, 2019), VSR (Liu et al., 2022), SpatialSense (Yang et al., 2019), MMBench (Liu et al., 2023d), etc. The reason is that this work aims to provide a carefully curated benchmark for spatial relationship understanding evaluation in a controlled setting to abstract away intervening factors that make the evaluations noisy.

Limited Spatial Prepositions: Following the What'sUp dataset, our benchmark is also confined to the primitive spatial clauses of *on*, *under*, *behind*, *in front of*, *to the left of*, *to the right of*, *below* and *above*, when having two objects involved in the caption, and, *on the top*, *on the bottom*, *on the left* and *on the right* when having only one object in the caption, like in COCO-Spatial-One and GQA-Spatial-One.

Lack of Robustness in MC Prompting: In addition to the similar findings of MC noisiness in LLMs that we discussed earlier, we hypothesize that the higher degree of variance in multiple-choice results in the last two subsets (COCO and GQA), which is more significant in the smaller models, could be due to the language domain distribution shift. Most of the LLMs and MLLMs are being trained and evaluated with 4 options in the multiple-choice settings. Conversely, in the last two subsets, we have two captions per image, which means we only provide options A and B to the model in the prompt instead of ABCD without any fine-tuning for this task or this specific type of prompting.

Intern-VL-1.5 Poor Grounding Observation: An unexpected, significant noisiness in the output of grounding/localization prompting of InternVL-1.5 model prevented us from analyzing and reporting the results for this model, which requires further investigation since a similar behavior has been observed through our interaction with the InternVL-1.5 demo, as well.

Depth Augmentation Nuances: The issue we noticed in the DAP experiment was the distraction the depth hint can cause in cases where multiple correct relationships hold in the image. For instance, object A can be *to the left of* object B, and also *in front of* object B, at the same time. So, in these ambiguous cases, incorporating depth could make the model's decision biased towards the *in front of* preposition, while the ground-truth might be *to the left of* in this case. Therefore, we believe that trying both prompts, with and w/o depth hint, would be helpful for disambiguation in such cases.

No Human Annotation: Due to the resource constraints, our extended benchmark relies on the pseudo-labels we generate using state-of-the-art, off-the-shelf models like GroundingDINO, SAM, and ZoeDepth. Future works could incorporate human inspection and labeling for further robustness in annotations.

References

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual Question Answering. In *International Conference on Computer Vision (ICCV)*.

Shariq Farooq Bhat, Reiner Birkl, Diana Wofk, Peter Wonka, and Matthias Müller. 2023. Zoedepth: Zero-shot transfer by combining relative and metric depth. *arXiv preprint arXiv:2302.12288*.

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

Drew A Hudson and Christopher D Manning. 2019. 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.

Amita Kamath, Jack Hessel, and Kai-Wei Chang. 2023. 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*, pages 9161–9175, Singapore. Association for Computational Linguistics.

Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. 2014. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 787–798.

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

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123:32–73.

Bo Li, Kaichen Zhang, Hao Zhang, Dong Guo, Renrui Zhang, Feng Li, Yuanhan Zhang, Ziwei Liu, and Chunyuan Li. 2024. Llava-next: Stronger llms supercharge multimodal capabilities in the wild.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR.

Fangyu Liu, Guy Emerson, and Nigel Collier. 2022. Visual spatial reasoning. *arXiv preprint arXiv:2205.00363*.

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

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

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023b. Visual instruction tuning. In *NeurIPS*.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. 2023c. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*.

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

Pouya Pezeshkpour and Estevam Hruschka. 2023. Large language models sensitivity to the order of options in multiple-choice questions. *arXiv preprint arXiv:2308.11483*.

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

Navid Rajabi and Jana Kosecka. 2023. Towards grounded visual spatial reasoning in multi-modal vision language models. *arXiv preprint arXiv:2308.09778*.

Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy Lillicrap, Jean-baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. 2022. Flava: A foundational language and vision alignment model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15638–15650.

Hao Tan and Mohit Bansal. 2019. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*.

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Qi Liu, Tianyu Liu, and Zhifang Sui. 2023. Large language models are not fair evaluators. *arXiv preprint arXiv:2305.17926*.

Xinpeng Wang, Chengzhi Hu, Bolei Ma, Paul Röttger, and Barbara Plank. 2024. Look at the text: Instruction-tuned language models are more robust multiple choice selectors than you think. *arXiv preprint arXiv:2404.08382*.

Mengge Xue, Zhenyu Hu, Meng Zhao, Liqun Liu, Kuo Liao, Shuang Li, Honglin Han, and Chengguo Yin. 2024. Strengthened symbol binding makes large language models reliable multiple-choice selectors.

Kaiyu Yang, Olga Russakovsky, and Jia Deng. 2019. Spatialsense: An adversarially crowdsourced benchmark for spatial relation recognition. In *International Conference on Computer Vision (ICCV)*.

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

Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. 2022. When and why vision-language models behave like bag-of-words models, and what to do about it? *arXiv preprint arXiv:2210.01936*.

Yan Zeng, Xinsong Zhang, and Hang Li. 2021. Multi-grained vision language pre-training: Aligning texts with visual concepts. *arXiv preprint arXiv:2111.08276*.

Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. 2023. Large language models are not robust multiple choice selectors. In *The Twelfth International Conference on Learning Representations*.

A Appendix

The appendix is organized as follows:

- Figure 4 demonstrates the biases of multiple-choice (MC) prompting.
- Figures 5 - 11 depict the distributions of objects occurring in the captions.
- Figure 12 shows sample failures in grounding small objects in Subset A.

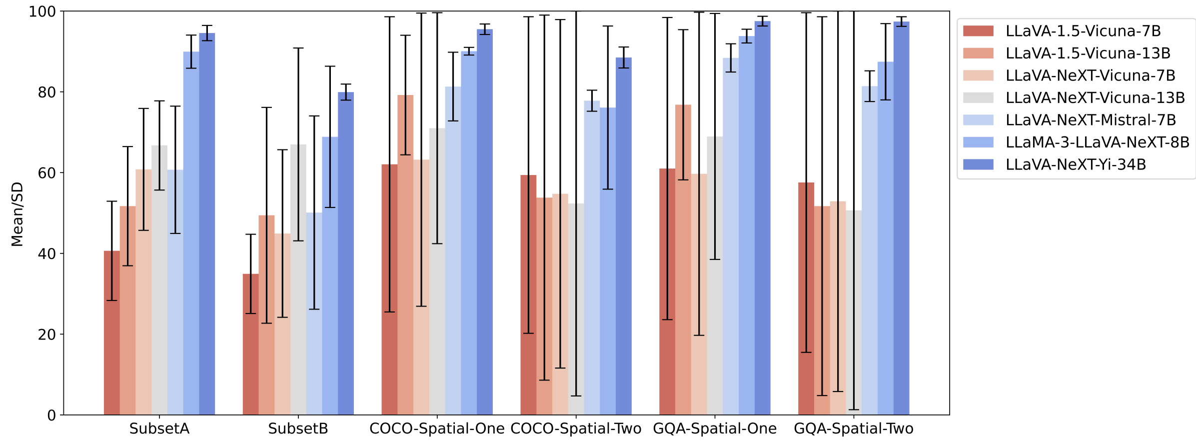


Figure 4: Sensitivity of the models to different permutations of choice order, in the multiple-choice (MC) experiment, which is more significant in the smaller models, and when having two choices of A and B instead of regular 4-choice of A, B, C, and D. LLAVA-NEXT-YI-34B demonstrates an excellent robustness against this issue.

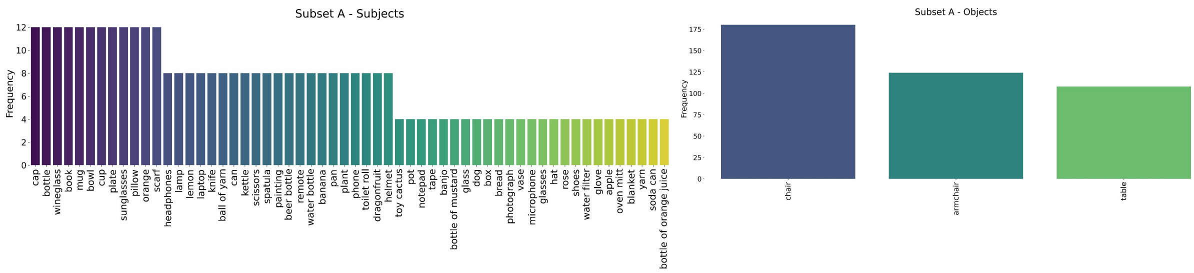


Figure 5: Subset A - subjects and objects distributions.

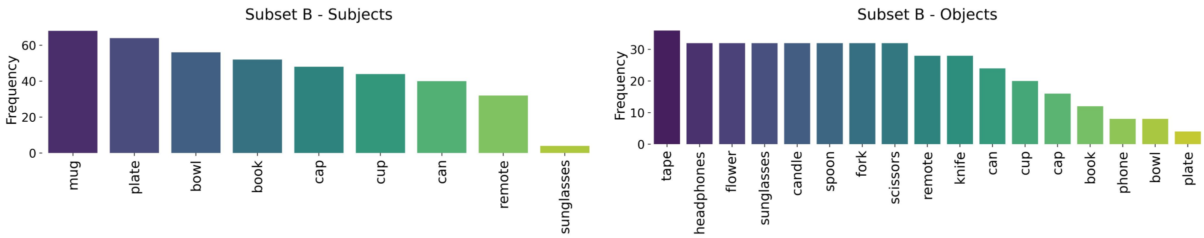


Figure 6: Subset B - subjects and objects distributions.

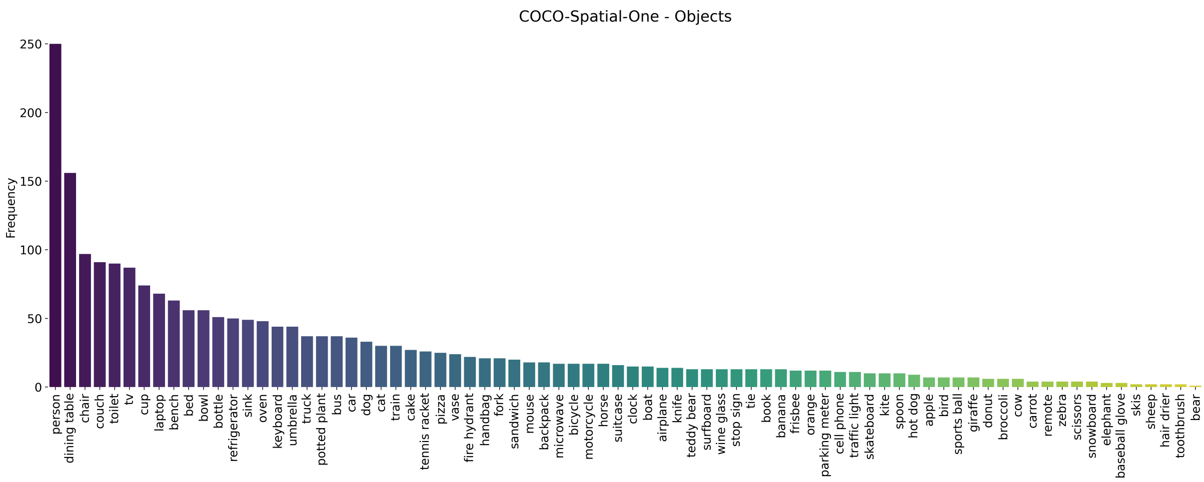


Figure 7: COCO Spatial One - objects distribution.

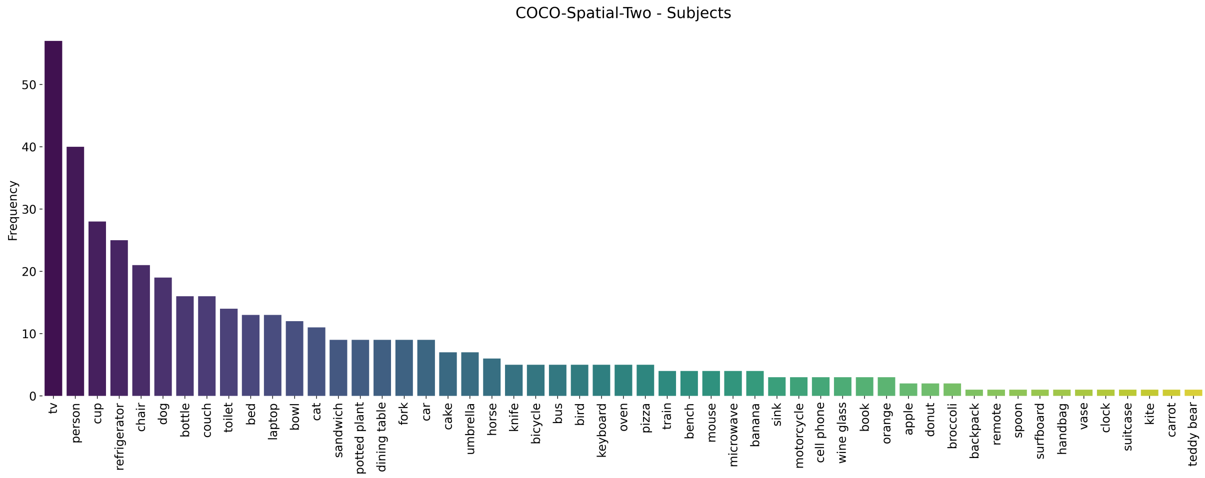


Figure 8: COCO Spatial Two - subjects distribution.

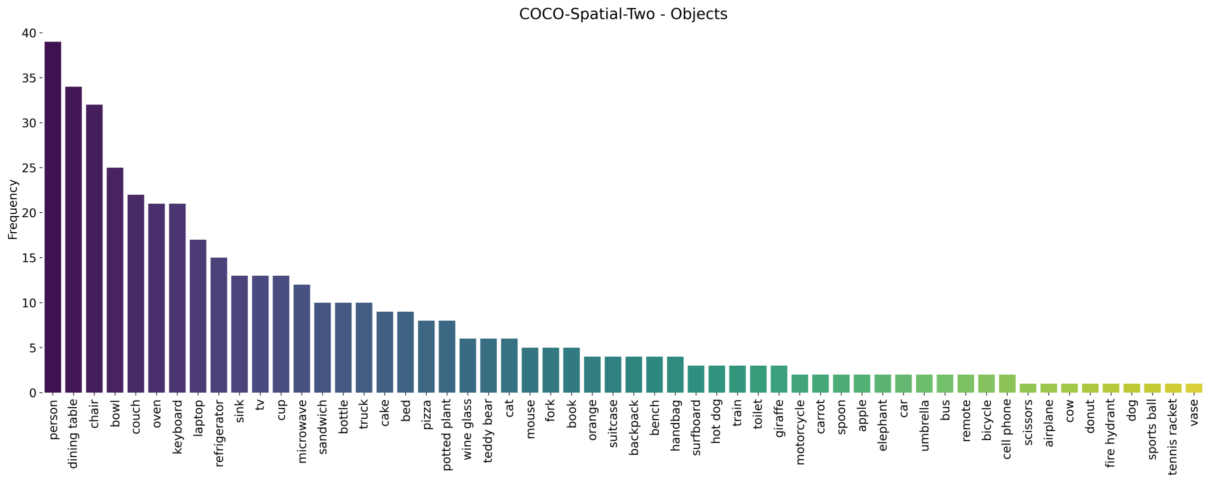


Figure 9: COCO Spatial Two - objects distribution.

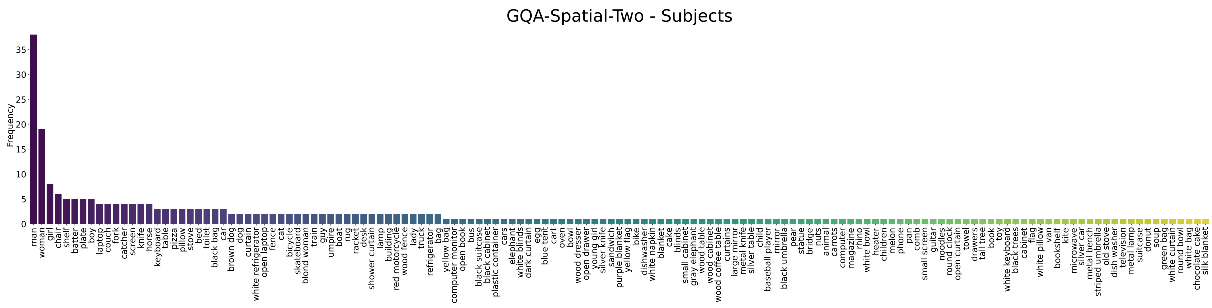


Figure 10: GQA Spatial Two - subjects distribution.

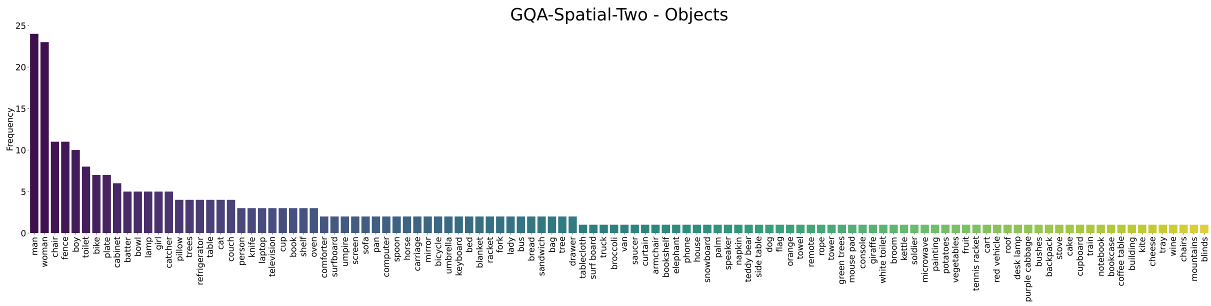
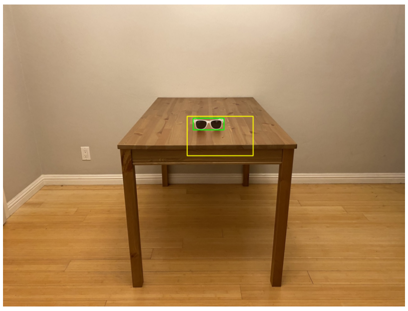


Figure 11: GQA Spatial Two - objects distribution.

(a) **toy cactus** on chair

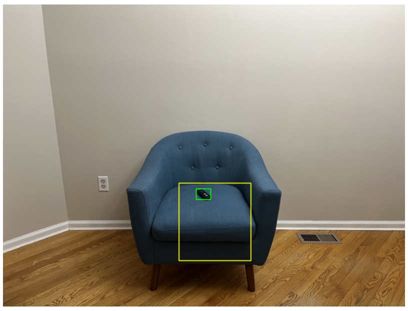
(b) **wineglass** under armchair

(c) **tape** under armchair



(d) **sunglasses** on table

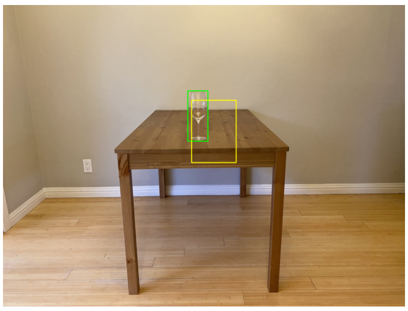
(e) **spatula** on chair



(f) **remote** on armchair

(g) **orange** right of armchair

(h) **ball of yarn** left of table



(i) **wineglass** on table

(j) **banjo** under armchair

Figure 12: Sample failures in small objects grounding (i.e., IoU < 0.5), which refers to the SUB column results of Subset A in Table 2. The pseudo-ground-truth bounding box, which is the GroudningDINO output, is indicated in green, and the output of LLAVA-NEXT-QWEN-1.5-110B, which is the best-performing MLLM in our grounding/localization experiment, is demonstrated in yellow.