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

Navid Rajabi
Department of Computer Science
George Mason University
nrajabi@gmu.edu

Jana Košecká
Department of Computer Science
George Mason University
kosecka@gmu.edu

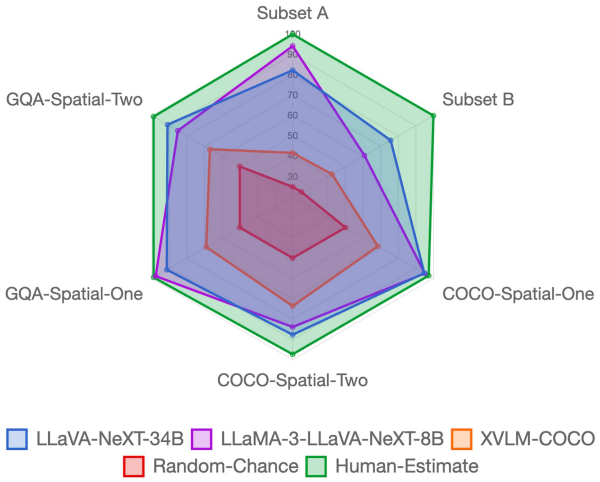


Figure 1: LLAMA-3-LLAVA-NEXT-8B achieves the overall accuracy of 86.1%, compared to 60.4% by XVLM-COCO, in What'sUp benchmark, reaching the best trade-off between accuracy and parameters size, since it performs only 1.1% lower than LLAVA-NEXT-34B, which has $\times$4.25 number of parameters.

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 recognizing and localizing objects of interest and determining 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 [5] dataset and propose a novel comprehensive evaluation for spatial relationship understanding that highlights the strengths and weaknesses of 9 Multimodal LLMs (MLLMs), in comparison with the 18 VLMs tested in What'sUp dataset. Our experiments encompass three classes of 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.

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 [23], CLIP [19], BLIP [10], with the focus on downstream tasks performances such as VQA [1], GQA [4], referring expressions [6], 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

NeurIPS 2024 Workshop on Compositional Learning: Perspectives, Methods, and Paths Forward.

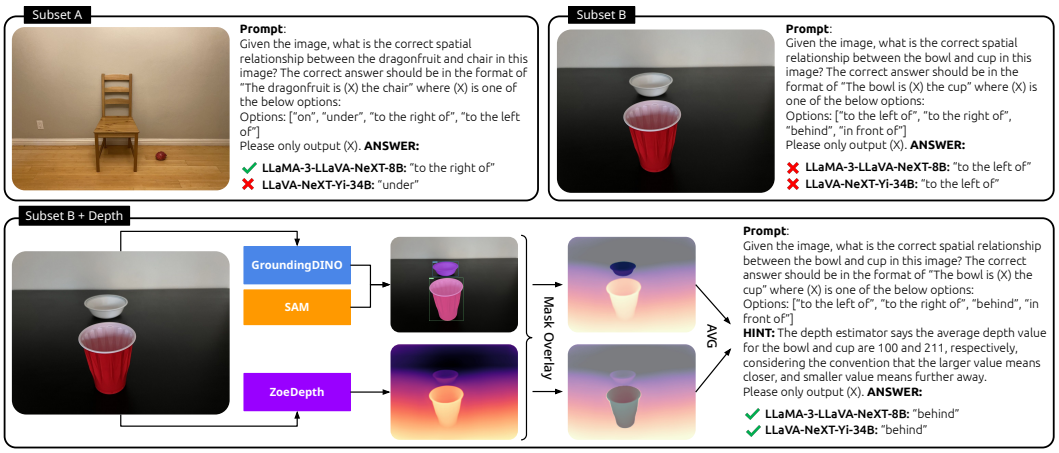


Figure 2: Our pipeline overview for spatial relationship understanding prompting, shown in the top two figures, and our depth-augmented prompting, shown in the bottom figure.

and their attributes[1]. To isolate these issues, authors in [5] 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 [5] 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 benchmarks like MMBench [17] 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: (1) Extended What'sUp spatial relationship dataset with depth, masks, and bounding box annotations, (2) Design of different prompting strategies through structured prompting for the evaluation of grounding and spatial reasoning, and (3) 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 [5] 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 [16] as an open-vocabulary object detector, Segment Anything (SAM) [7] for the object mask segmentation, and ZoeDepth [2] 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].

[1]VQA example question may be: "Is there a woman to the left of the person that is wearing a wetsuit?"
[2]All the code and data will be publicly available.

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 [19]	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 [30]	–	32.5	36.3	47.4	46.4	45.3	46.7	42.4
RoBERTaCLIP [5]	–	25.2	25.0	46.3	53.6	50.8	48.8	41.6
CoCa [29]	2.1B	29.4	29.4	48.1	45.2	45.0	49.1	41.0
XVLM 4M [31]	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 [10]	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 [11]	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 [22]	–	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 [5] results. The rest shows our LLaVA 1.5 , 1.6 , and Intern-VL-1.5 prompting results. Our best-performing is **bold**, 2nd-best underlined, and What'sUp best-performing *italic underlined*.

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 [5], we focus on the probing of open-source generative MLLMs like LLaVA and Intern-VL[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. We ran each prompt with 4

MODEL	SUBSET A SUB	SUBSET A OBJ	SUBSET B SUB	SUBSET B OBJ	COCO-SPATIAL ONE-OBJ	COCO-SPATIAL SUB	COCO-SPATIAL OBJ	GQA-SPATIAL ONE-OBJ	GQA-SPATIAL SUB	GQA-SPATIAL OBJ	AVG 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	87.1	86.8	85.7	69.2	85.9	81.8	74.58
LLAVA-NEXT-VICUNA-13B	15.3	84.0	95.3	67.6	87.1	**90.2**	83.9	69.6	85.9	80.4	75.93
LLAMA-3-LLAVA-NEXT-8B	19.2	99.3	96.6	73.8	85.7	87.5	83.2	69.0	84.5	80.4	77.92
LLAVA-NEXT-YI-34B	21.1	**100.0**	97.8	78.9	83.7	85.7	81.4	70.0	**88.0**	83.5	79.01
LLAVA-NEXT-QWEN1.5-110B	**29.4**	98.5	**98.8**	**80.1**	**88.7**	88.2	**86.4**	**74.9**	86.9	**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.

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 [17]. As opposed to the CircularEval,

[3]Intern-VL is the leading model in MMBench.

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 3 in the Appendix for details). This phenomenon also corroborates the findings of multiple recent studies in LLMs [32, 18, 24, 27, 25]. According to this observation, we opted for TG prompting, accompanied by the CircularEval methodology, inspired by Gemini 1.5 Pro [21]. See Table 1 for the TG prompting results, where rows in section 1 and 2 come from the What'sUp benchmark [5], section 3 refers to LLaVA-1.5 models [14], section 4 to the LLaVA-NeXT models [9, 15], and section 5 to the InternVL-1.5 results [3].

Grounding/Localization Evaluation. This experiment aims to measure the MLLMs grounding ability of the objects mentioned in the captions. Recent studies like [20] on Visual Spatial Reasoning (VSR) benchmark [13] 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.

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.

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.

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 in reasoning while grounding is still going up by scaling. Since the lowest grounding results have been achieved in localizing the subjects in Subset A, we also ran an additional Causal Effect analysis using linear regression to measure the causality between the grounding accuracy, as the independent variable, and the reasoning accuracy, as the dependent variable. As a result, we could reject the null hypothesis that "There is no causal effect between grounding and reasoning" due to the regression coefficient of β = 6.025 and the ρ-value of 0.007, which is significantly lower than 0.05, shown in Figure 4.

4 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.

References

[1] Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C. L., and Parikh, D. (2015). VQA: Visual Question Answering. In *International Conference on Computer Vision (ICCV)*.

[2] Bhat, S. F., Birkl, R., Wofk, D., Wonka, P., and Müller, M. (2023). Zoedepth: Zero-shot transfer by combining relative and metric depth. *arXiv preprint arXiv:2302.12288*.

[3] Chen, Z., Wang, W., Tian, H., Ye, S., Gao, Z., Cui, E., Tong, W., Hu, K., Luo, J., Ma, Z., *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*.

[4] Hudson, D. A. and Manning, C. D. (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.

[5] Kamath, A., Hessel, J., and Chang, K.-W. (2023). What's "up" with vision-language models? investigating their struggle with spatial reasoning. In H. Bouamor, J. Pino, and K. Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9161–9175, Singapore. Association for Computational Linguistics.

[6] Kazemzadeh, S., Ordonez, V., Matten, M., and Berg, T. (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.

[7] Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W.-Y., *et al.* (2023). Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026.

[8] Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.-J., Shamma, D. A., *et al.* (2017). Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, **123**, 32–73.

[9] Li, B., Zhang, K., Zhang, H., Guo, D., Zhang, R., Li, F., Zhang, Y., Liu, Z., and Li, C. (2024). Llava-next: Stronger llms supercharge multimodal capabilities in the wild.

[10] Li, J., Li, D., Xiong, C., and Hoi, S. (2022). Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR.

[11] Li, J., Li, D., Savarese, S., and Hoi, S. (2023). Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*.

[12] Liu, F., Emerson, G., and Collier, N. (2022). Visual spatial reasoning. *arXiv preprint arXiv:2205.00363*.

[13] Liu, F., Emerson, G., and Collier, N. (2023a). Visual spatial reasoning. *Transactions of the Association for Computational Linguistics*, **11**, 635–651.

[14] Liu, H., Li, C., Wu, Q., and Lee, Y. J. (2023b). Visual instruction tuning. In *NeurIPS*.

[15] Liu, H., Li, C., Li, Y., Li, B., Zhang, Y., Shen, S., and Lee, Y. J. (2024). Llava-next: Improved reasoning, ocr, and world knowledge.

[16] Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Li, C., Yang, J., Su, H., Zhu, J., *et al.* (2023c). Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*.

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

[18] Pezeshkpour, P. and Hruschka, E. (2023). Large language models sensitivity to the order of options in multiple-choice questions. *arXiv preprint arXiv:2308.11483*.

[19] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., *et al.* (2021). Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR.

[20] Rajabi, N. and Kosecka, J. (2023). Towards grounded visual spatial reasoning in multi-modal vision language models. *arXiv preprint arXiv:2308.09778*.

[21] Reid, M., Savinov, N., Teplyashin, D., Lepikhin, D., Lillicrap, T., Alayrac, J.-b., Soricut, R., Lazaridou, A., Firat, O., Schrittwieser, J., *et al.* (2024). Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

[22] Singh, A., Hu, R., Goswami, V., Couairon, G., Galuba, W., Rohrbach, M., and Kiela, D. (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.

[23] Tan, H. and Bansal, M. (2019). Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*.

[24] Wang, P., Li, L., Chen, L., Cai, Z., Zhu, D., Lin, B., Cao, Y., Liu, Q., Liu, T., and Sui, Z. (2023). Large language models are not fair evaluators. *arXiv preprint arXiv:2305.17926*.

[25] Wang, X., Hu, C., Ma, B., Röttger, P., and Plank, B. (2024). Look at the text: Instruction-tuned language models are more robust multiple choice selectors than you think. *arXiv preprint arXiv:2404.08382*.

[26] Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., *et al.* (2019). Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

[27] Xue, M., Hu, Z., Zhao, M., Liu, L., Liao, K., Li, S., Han, H., and Yin, C. (2024). Strengthened symbol binding makes large language models reliable multiple-choice selectors.

[28] Yang, K., Russakovsky, O., and Deng, J. (2019). Spatialsense: An adversarially crowdsourced benchmark for spatial relation recognition. In *International Conference on Computer Vision (ICCV)*.

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

[30] Yuksekgonul, M., Bianchi, F., Kalluri, P., Jurafsky, D., and Zou, J. (2022). When and why vision-language models behave like bag-of-words models, and what to do about it? *arXiv preprint arXiv:2210.01936*.

[31] Zeng, Y., Zhang, X., and Li, H. (2021). Multi-grained vision language pre-training: Aligning texts with visual concepts. *arXiv preprint arXiv:2111.08276*.

[32] Zheng, C., Zhou, H., Meng, F., Zhou, J., and Huang, M. (2023). Large language models are not robust multiple choice selectors. In *The Twelfth International Conference on Learning Representations*.

A Appendix / supplemental material

A.1 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 [8], GQA [4], VSR [12], SpatialSense [28], MMBench [17], 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.

A.2 Computational Resources

For LLaVA-1.5 (7B and 13B) and LLaVA-1.6/LLaVA-NeXT models up to the 13B versions, we used one 40 GB NVIDIA A100. For LLaVA-Yi-34B, and Intern-VL-Chat-1.5 26B, we used one 80 GB NVIDIA A100. For LLaVA-NeXT-Qwen-1.5 110B, we used four 80 GB NVIDIA A100 GPUs. In most cases, except the Intern-VL model, we used the Huggingface [26] library setup and checkpoints.

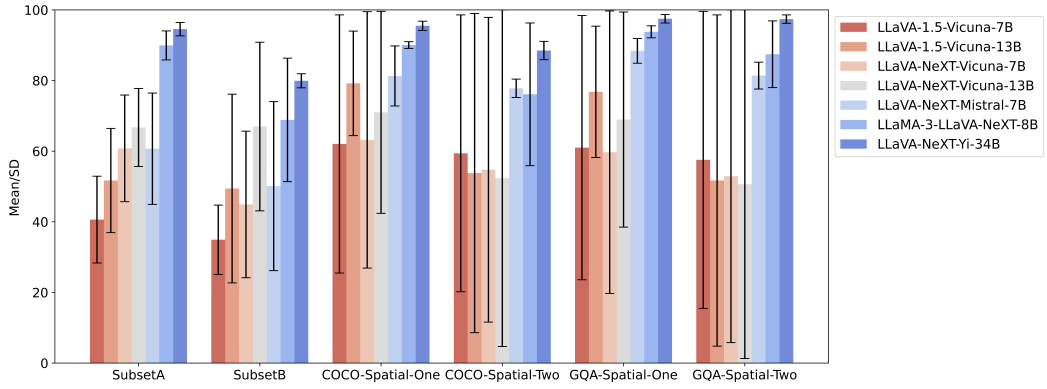


Figure 3: Sensitivity of the models to different permutations of choice order, in the multiple-choice (MC) experiment, which is more significant in the smaller models, and when having two choices of A and B instead of regular 4-choice of A, B, C, and D. LLaVA-NeXT-Yi-34B demonstrates an excellent robustness against this issue.

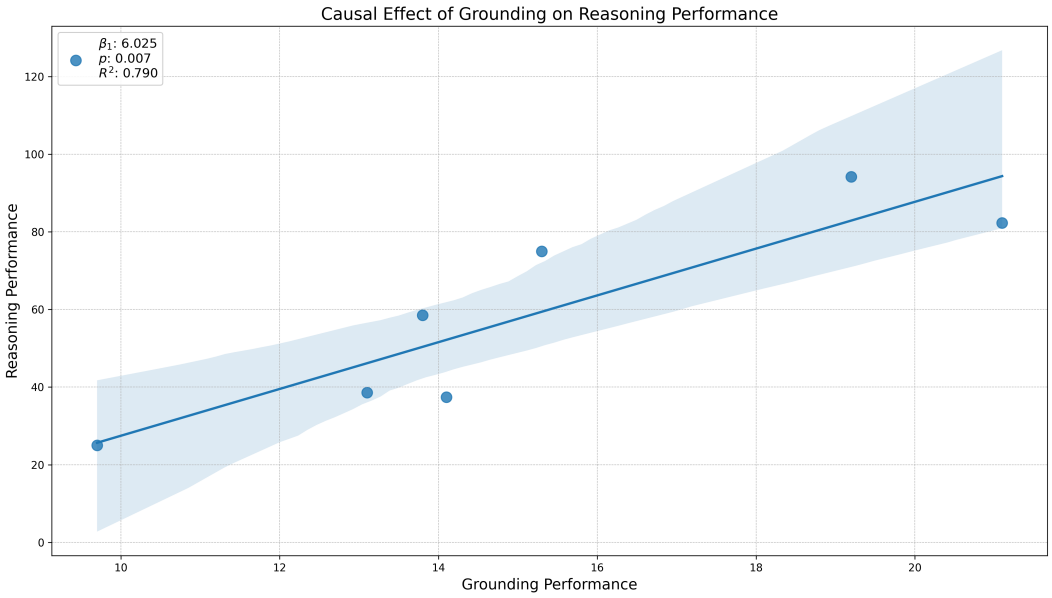
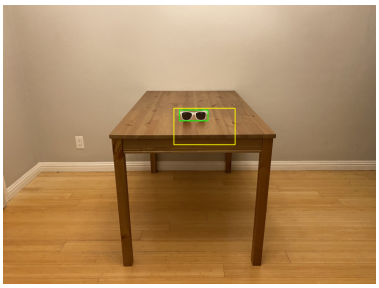


Figure 4: Causal effect analysis between grounding and reasoning accuracy, for the Subset A subjects that was the most difficult setting for the models for localization.

(a) **toy cactus** on chair

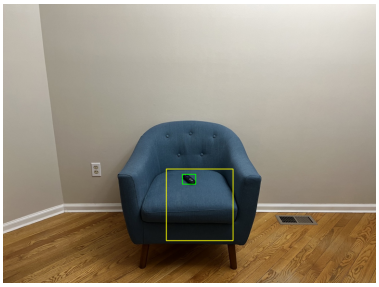
(b) **wineglass** under armchair

(c) **tape** under armchair



(d) **sunglasses** on table

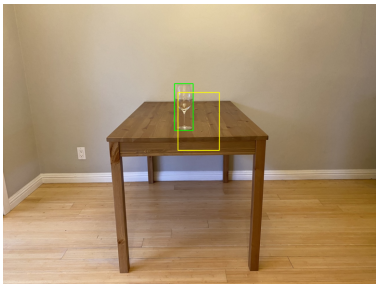
(e) **spatula** on chair



(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 5: 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.