Referring Expressions as a Lens into Spatial Language Grounding in Vision-Language Models

Akshar Tumu
UC San Diego /
La Jolla, CA, USA
atumu@ucsd.edu

Varad Shinde
IIT Kanpur /
Kanpur, Uttar Pradesh, India
varadshinde@iitbhilai.ac.in

Parisa Kordjamshidi
Michigan State University /
East Lansing, MI, USA
kordjams@msu.edu

Abstract

Spatial Reasoning is an important component of human cognition and is an area in which the latest Vision-language models (VLMs) show signs of difficulty. The current analysis papers often use image captioning tasks and visual question answering. In this work, we propose using the Referring Expression Comprehension task instead as a platform for the evaluation of spatial reasoning by VLMs. This platform provides the opportunity for a deeper analysis of spatial comprehension and grounding abilities when there is 1) ambiguity in object detection, 2) complex spatial expressions with a longer sentence structure and multiple spatial relations, and 3) expressions with negation ('not'). In our analysis, we use task-specific architectures as well as large VLMs and highlight their strengths and weaknesses in dealing with these specific situations. While all these models face challenges with the task at hand, the relative behaviors depend on the underlying models and the specific categories of spatial semantics (topological, directional, proximal, etc.). Our results highlight these challenges and behaviors and provide insight into research gaps and future directions.

1 Introduction

Vision-language model (VLM) research has boomed in the recent past, owing to the enhanced user interaction and accessibility they provide. Models such as GPT 4o[1], LLaVA (Liu et al., 2024), Google Gemini (Team et al., 2023) have become adept at solving vision-language tasks such as Visual Question Answering (VQA), Image Captioning, and more. However, VLMs still lack human-level 'Spatial Reasoning' capabilities (Liu et al., 2023a; Kamath et al., 2023). Spatial reasoning involves comprehending relations that depict the absolute/relative position or orientation of an object, such as 'left', 'above', or 'near'. Inaccurate spatial reasoning by VLMs can lead to serious consequences in embodied AI domains such as autonomous driving and surgical robotics. A focused analysis of VLMs' spatial reasoning capabilities can help identify potential reasoning issues.

Most of the previous works confine their analysis to testing which models work well for spatial relations. We go further to analyze the comparative performance of these models for spatial categories that represent different orientational and positional relations between objects. A novel aspect of our work is the analysis of the effect of varying spatial composition (number of spatial relations) in the expressions on the performance of the models.

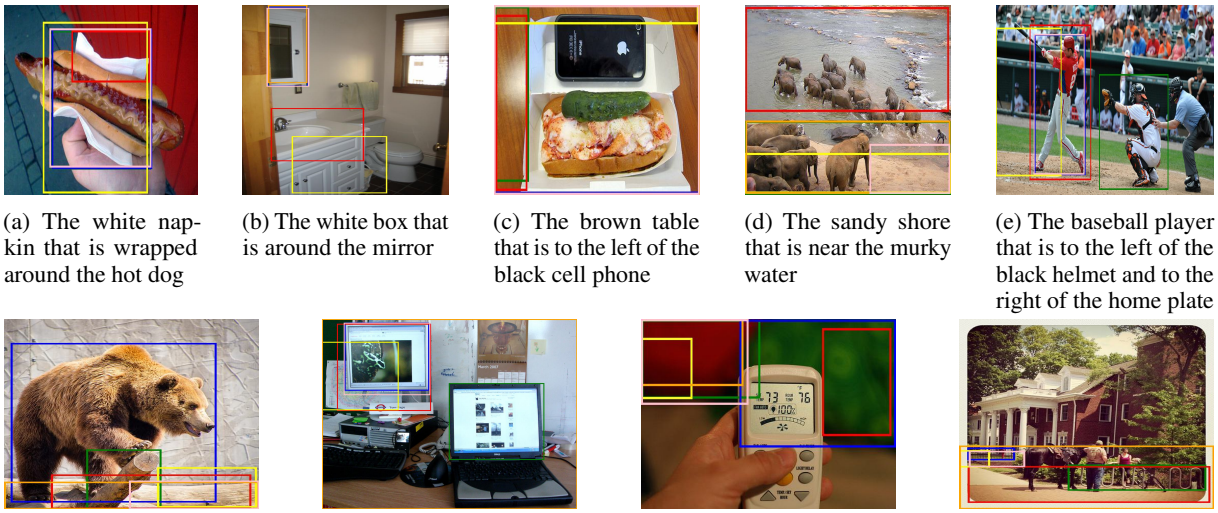
Previous works focused on spatial analysis with image captioning-related tasks, thus failing to locate the source of error in the presence of visual and linguistic ambiguity. To avoid this, we adopt the Referring Expression Comprehension (REC) task (Qiao et al., 2020) where the models output bounding boxes around the target entity based on a natural language expression, the analysis of which could reveal the parts of the input that the models fail to comprehend. Comprehension accuracy (or simply, accuracy) is a common metric for this task; it captures how often a model correctly outputs the bounding box around the target entity.

We test four popular VLMs - LLaVA (Liu et al., 2024), Grounding DINO (GDINO) (Liu et al., 2023b), DeepSeek-VL2 (Wu et al., 2024), and Qwen2.5-VL (Bai et al., 2025). We also include 'MGA-Net' (Zheng et al., 2020), a model specifically designed for the REC task. These models offer diversity in the evaluation as they cover different architectural elements, training strategies, and input formats. We further compare these models with an object detector baseline to test if the images are truly complex and require elaborate referring expressions to ground the correct object.

Some of our important findings are as follows:
(1) Referring expressions that include spatial re-

[1] https://openai.com/index/hello-gpt-4o/

Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 3442–3455
December 20-24, 2025 ©2025 Association for Computational Linguistics



(a) The white napkin that is wrapped around the hot dog

(b) The white box that is around the mirror

(c) The brown table that is to the left of the black cell phone

(d) The sandy shore that is near the murky water

(e) The baseball player that is to the left of the black helmet and to the right of the home plate

(f) The large branch that is to the right of the log that is behind the large bear

(g) The black monitor that is to the left of the keyboard or on the desk

(h) The blanket that is not green and that is not on the bed

(i) The fence that is not black and that is not to the left of the man

Figure 1: Figures for qualitative analysis. Bounding box legend - Red: MGA-Net, Blue: GDINO, Yellow: LLaVA, Orange: DeepSeek-VL2, Pink: Qwen2.5-VL, Green: Ground-truth

lations, in addition to object attributes, result in higher accuracy on the REC task compared to expressions with only attributes. **(2)** Increasing the spatial complexity (no. of spatial relations) of an expression affects the performance of the VLMs, but models with explicit compositional learning components maintain the performance. **(3)** Expressions involving dynamic spatial relations yield low average accuracy across models, indicating the difficulty in modelling these relations. **(4)** The task-specific trained models achieve higher accuracy for expressions with geometric spatial relations (e.g., left of, right of) while the VLMs show relatively better accuracy for expressions having ambiguous relations such as proximity. **(5)** The models fail to recognize negated spatial relations in referring expressions in multiple instances, though the extent of this failure varies across models.

2 Related Work

Previous works have conducted a broad analysis on the ability of VLMs to perform multimodal perception and reasoning tasks, such as Spatial Reasoning, Multimodal conversation, etc. Many comprehensive real-world benchmarks have been introduced to test multiple VLM capabilities. (Liu et al., 2024; Tian et al., 2024; Liu et al., 2023c; Fu et al., 2025).

Some works (Subramanian et al., 2022; Rösch and Libovický, 2023) focus solely on spatial analysis of VLMs. Wang et al., 2024 go a step further to analyze the role of each modality in spatial reasoning. However, these works do not analyze the factors that affect the spatial reasoning ability of the VLMs. Another class of works performs a category-wise analysis of spatial relations, either based on their spatial properties (Liu et al., 2023a; Gokhale et al., 2022) or their linguistic properties and complexity (Kuhnle et al., 2018). In contrast, Kamath et al., 2023 analyze the effects of spatial biases in the datasets for the REC task performance.

Task Complexity and Interpretability. The above-mentioned works use image-caption agreement as their evaluation task. Due to the inherent limitations of this task, these works simplified the expressions to have only 2 objects and 1 spatial relation. To improve the interpretability of model output, synthetic datasets have been used instead of real-world images (Subramanian et al., 2022; Lewis et al., 2022). However, it simplifies the problem due to bounded expressivity (a limited set of attributes and spatial relations). On the other hand, REC models output bounding boxes around the target objects. Analyzing the characteristics of these boxes helps identify the parts of the input that the models fail to process. This enables comparative analysis of expressions with 0, 1, or more spatial relations, a unique feature of our work. The REC task also enables us to test the models over images of different visual complexities (single or multiple instances of objects in an image).

Dataset	Object Categories	Image Source	Avg. Length of Expression	Avg. No. of Objects per Image	No. of Spatial Relations
RefCOCO	80	Real-world	3.6	10.6	59
RefCOCOg	80	Real-world	8.4	8.2	72
CLEVR-Ref+	3	Synthetic	22.4	6.5	4
CopsRef	508	Real-world	14.4	17.4	51

Table 1: Statistics of Popular Referring Expression Comprehension Datasets. For the last column, the relation types are taken from various resources explained in Section 2 and (Marchi Fagundes et al., 2021), in addition to the relations in Table 2.

Category	Number	Spatial Relations
Absolute	56	on the right, on the left, in the middle, in the center, from the right, from the left
Adjacency	14	attached, against, on the side, on the back, on the front, on the edge
Directional	29	falling off, along, through, across, down, up, hanging from, coming from, around
Orientation	0	facing
Projective	2361	on top of, beneath, beside, behind, to the left, to the right, under, above, in front of, over, below, underneath
Proximity	217	by, close to, near
Topological	1054	connected, contain, with, surrounding, surrounded by, inside, between, touching, out of, at, in, on
Unallocated	56	next to, enclosing

Table 2: Category-wise relation split and number of referring expressions in the CopsRef (Chen et al., 2020) test set with 1 spatial relation in each category. While there are no referring expressions with only one relation from the orientation category, these relations co-occur with relations from other categories in some expressions.

Related Tasks - Embodied Settings. Beyond the REC task, spatial reasoning work also spans embodied settings. Spatial affordance prediction involves localizing precise action points in the image based on a spatial language instruction(Yuan et al., 2024). Benchmarks such as Room-to-Room (R2R) (Anderson et al., 2018) involve agents following natural-language route descriptions through real buildings, requiring robust spatial reasoning over landmarks and paths. Embodied Spatial Analysis, which focuses on the effects of different perspectives and non-verbal cues on the spatial reasoning capabilities of VLMs (Islam et al., 2022, 2023).

3 Dataset

Table 1 shows the key characteristics of some popular REC datasets. We chose CopsRef (Chen et al., 2020) because among real-world datasets, it has the longest referring expressions, which go beyond describing the simple, distinctive properties of the objects. CopsRef is also a highly spatial dataset, as 90% of expressions consist of spatial relations. Examples of such referring expressions and the corresponding images are given in Figure 1. The total no. of expressions in the train/validation/test splits are 119,603/16,524/12,586 respectively. Table 2 shows the category-wise split of the 51 spatial relations we identified in the CopsRef test dataset. For explanations of each category, refer to Appendix A.

Category	No. of relations	No. of expressions
None	0	1202
One	1	3787
Two-chained	2	1324
Two-and	2	3890
Two-or	2	2203
Three	3	180

Table 3: Frequency of occurrence of spatial relations in referring expressions.

Table 3 shows the number of expressions having 0, 1, 2, and 3 spatial relations. For expressions with 2 spatial relations, we have introduced three categories. The first category, 'Two-chained', includes expressions where the spatial clauses are chained sequentially. The second category, 'Two-and', contains expressions where the referred object satisfies both spatial clauses. Finally, 'Two-or' consists of expressions where the referred object satisfies either of the two spatial clauses. Figures 1e-1g illustrate examples of the three categories.

4 Approach

In our analysis, we consider spatial reasoning to include both basic grounding (locating the right entities) and spatial relations (reasoning about where entities are relative to one another). These aspects are inseparable in real scenes as attributes and ob-

ject properties (size, shape, rigidity, affordances) determine which relations are even plausible. Eg: a rock submerges in water, while a feather tends to float on the surface of water. Decoupling them might yield an unrealistic toy setup. We use REC as the analysis medium because it exposes failures in either component. We therefore seek to the following research questions:
RQ1. Which spatial relation categories result in low accuracy for REC models? **RQ2.** How do different model characteristics/architectures influence the REC task accuracy for certain spatial relation categories compared to the others? **RQ3.** Does the inclusion of spatial relations increase or decrease the accuracy of REC models? **RQ4.** How does the number of spatial relations in the expressions affect the accuracy across different types of models? **RQ5.** Do the REC models accurately recognize negated spatial relations in expressions?

To answer these questions, we outline our research methodology and the designed experiments in this section:

4.1 Models Description

MGA-Net. (Zheng et al., 2020) It is an REC task-specific model with a compositional architecture designed to handle complex expressions. The model uses soft attention to decompose the expression and builds a relational graph among objects using the connecting spatial relations and attributes. A Gated Graph Neural Network then performs multi-step reasoning over this graph. We use Faster R-CNN (Ren et al., 2016) to generate object proposals and extract their features using a pre-trained ResNet-101. The model is designed to use the individual bounding box features instead of the image as a whole.

Grounding DINO (GDINO). (Liu et al., 2023b) It is an open-set object detector VLM with both vision and language backbones, whose outputs are fused at multiple levels. Grounded pre-training with contrastive loss makes it well-suited for the REC task. We use the Swin-B vision backbone and the CLIP-text encoder (Radford et al., 2021) for the language backbone. We filter all bounding box detections for an expression using their output labels to see which detections match the target entity. Then we select the detection with the highest confidence score.

LLaVA. (Liu et al., 2024) It is a general-purpose VLM that connects an open-set vision encoder from CLIP with a language decoder. The model is trained end-to-end, which involves general visual instruction tuning for aligning the vision and language modalities.

DeepSeek-VL2. (Wu et al., 2024) It is a popular vision-language model with a dynamic tiling-based vision encoder (SigLIP) and a Mixture-of-Experts (MoE) language model, whose features are aligned using a vision-language adaptor. The model has been trained for grounded conversation (return the bounding box for an expression) using popular REC datasets like RefCOCO, which feature simple expressions with limited relational complexity. We employ the DeepSeek-VL2 Tiny variant of the model to suit the available computational resouces.

Qwen2.5-VL. (Bai et al., 2025) It is a multimodal model built over a Vision Transformer (ViT) encoder having native-resolution support to preserve spatial information and a Qwen2.5 LLM decoder. An MLP-based vision-language merger aligns these two modalities. Qwen2.5-VL is supervised using both point and bounding box annotations for grounding tasks. The model is trained on diverse referring expressions with over 10,000 real and synthetic object categories.

OWL-ViT. (Minderer et al., 2022) It is an object detector baseline that only takes the target object's label as the input instead of the referring expression. It is an open-set object detector, suitable for the CopsRef dataset because its expressions include entities from the Visual Genome (Krishna et al., 2017) Scene Graphs, some of which are absent in datasets used to train famous closed-set detectors like YOLO (Redmon, 2016). It also has a simple architecture with a Vision transformer and CLIP for zero-shot image–label alignment, making it an ideal baseline.

4.2 Evaluation Dataset Splits

We create the following dataset test splits for evaluation and answering the earlier mentioned research questions, RQ1-RQ5.

Fine-grained Spatial Relations Split. In the test dataset, we split the expressions with 1 spatial relation using the categories shown in Table 2. Using the categories from Table 3, we split the remaining expressions based on the number of spatial relations they contain. Then, we rank the models based

on their accuracy for each category. To statistically find the correlation between the models' performance across categories, we employ the Kendall Tau Independence Test. The details about the test can be found in Appendix B.

Visual Complexity Split. To observe the effect of visual complexity on model performance, we split the test dataset into two parts. The first part has images that have multiple instances of one or more objects mentioned in the referring expressions. The second part has images with at most one instance of every object mentioned in the expression. We perform this splitting by first collecting the entities in each expression using spaCy[2] and then employing GDINO to find the number of instances in the image for each of the collected entities.

Negation Analysis Split. In our analysis, we found that models have difficulties in grounding spatial expressions with negations. Therefore, we created a test split for a more accurate evaluation and a deeper analysis of negated spatial expressions. We collected expressions that include the keyword 'not' and divided them into two sets according to the number of occurring negations (1 or 2). Then, we collected those expressions for which all the models give an IoU of less than 0.5. For each expression, we perform a qualitative analysis to verify whether the errors are due to misinterpreting the negations or conflation of other errors. We limit our analysis to the results from the first run of the models to facilitate the instance-wise analysis.

5 Results

Hardware. For inference of GDINO, LLaVA, OWLViT, and training of MGA-Net, we use the Quadro RTX 6000. Due to the heavy computational (GPU) requirements of DeepSeek-VL2 Tiny and Qwen2.5-VL, we use NVIDIA A100-SXM4 for their inference.

Evaluation. We evaluate the models using the Intersection over Union (IoU) metric. Following previous works (Yu et al., 2018; Chen et al., 2024), we consider the output as a correct comprehension if the IoU is greater than 0.5. We calculate the accuracy as the fraction of data points that have an IoU >0.5. We run each model three times (both training and testing for MGA-Net, and inference for the other models) to ensure the statistical significance of the evaluation.

[2] https://spacy.io/

5.1 Evaluation on Referring Expressions

Model	Accuracy (%)
MGA-Net	62.92 ± 0.11
GDINO	70.93 ± 0.01
LLaVA	34.96 ± 0.03
DeepSeek-VL2	30.07 ± 0
Qwen2.5-VL	67.00 ± 0
OWL-ViT	56.34 ± 0

Table 4: Comprehension Accuracies

We report the average accuracy and standard deviation across the three runs for each model in Table 4. Since we retrain MGA-Net for each run, the model predictions slightly change every time, resulting in the highest standard deviation among all models. VLMs and the baseline have zero or near-zero standard deviation since we test them zero-shot. This also follows for the future result tables.

GDINO performs the best, followed by Qwen and MGA-Net. LLaVA and DeepSeek perform worse than the baseline. A possible reason is that they don't possess a bounding box regression architecture. Another reason could be that while LLaVA doesn't have visual grounding instructions during pre-training, DeepSeek's training referring expressions often lack spatial complexity and use simple, non-relational phrases.

5.2 Evaluation on Fine-grained Relations

Table 5 shows a few general trends in results. The top 3-4 categories which each model performs the best for are categories with a single spatial relation. Among those, all the models perform well for the Topological and Absolute categories.

To answer **RQ1**, we observed that among all categories with a single spatial relation, the average accuracy across models is lowest for the Directional category expressions. A possible reason is that the spatial configurations of the involved objects are dynamic, as they vary from image to image for the same spatial relation. This makes it difficult for the models to learn common patterns for recognizing these relations, resulting in low accuracy.

5.3 Impact of Multiple Spatial Relations

Table 6 shows the Kendall Tau correlation values for all pairs of models. From the Kendall Tau Independence test, we observed that while all the VLMs are correlated, MGA-Net is not correlated with any

Category	MGA-Net		GDINO		LLaVA		DeepSeek-VL2		Qwen2.5-VL	
	Acc (%)	Rank	Acc (%)	Rank	Acc (%)	Rank	Acc (%)	Rank	Acc (%)	Rank
Absolute	70.24 ± 2.22	1	82.14 ± 0	2	44.64 ± 0	4	47.37 ± 0	3	80.70 ± 0	1
Adjacency	52.38 ± 3.37	12	78.57 ± 0	4	50.00 ± 0	1	71.43 ± 0	1	71.43 ± 0	4
Directional	52.87 ± 3.25	11	65.52 ± 0	12	27.59 ± 0	12	41.40 ± 0	5	70.69 ± 0	5
Projective	64.07 ± 0.08	4	69.12 ± 0	8	36.19 ± 0.08	6	34.57 ± 0	7	70.52 ± 0	6
Proximity	62.83 ± 0.22	8	80.65 ± 0	3	46.84 ± 0.22	3	45.16 ± 0	4	71.89 ± 0	3
Topological	67.32 ± 0.49	2	83.02 ± 0	1	48.51 ± 0.09	2	59.87 ± 0	2	72.91 ± 0	2
Unallocated	63.09 ± 0.84	6	75.00 ± 0	5	35.71 ± 0	7	30.36 ± 0	9	68.75 ± 0	7
None	62.39 ± 0.58	9	73.88 ± 0	6	42.89 ± 0.17	5	41.10 ± 0	6	66.14 ± 0	8
Two-chained	63.21 ± 0.22	5	70.67 ± 0.03	7	30.82 ± 0	9	18.13 ± 0	10	65.78 ± 0	9
Two-and	62.98 ± 0.09	7	68.45 ± 0.01	10	31.57 ± 0.07	8	31.45 ± 0	8	65.69 ± 0	10
Two-or	59.39 ± 0.21	10	68.97 ± 0.01	9	30.26 ± 0.04	10	12.68 ± 0	11	63.20 ± 0	11
Three	65.18 ± 2.1	3	67.78 ± 0	11	30.19 ± 0.26	11	11.11 ± 0	12	62.78 ± 0	12

Table 5: Category-wise accuracy and ranking

	MGA-Net	GDINO	LLaVA	Deep-Seek	Qwen
MGA-Net	1.00	0.18	0.09	-0.12	0.11
GDINO	0.18	1.00	0.73	0.52	0.63
LLaVA	0.09	0.73	1.00	0.73	0.53
DeepSeek	-0.12	0.52	0.73	1.00	0.75
Qwen	0.11	0.63	0.53	0.75	1.00

Table 6: Kendall Tau Independence Test results for category-wise ranks. DeepSeek model's version is VL2 and Qwen model's version is 2.5-VL.

of them. We study the reasons behind MGA-Net and VLMs differing in category-wise performance.

Among spatial categories of MGA-Net and VLMs, the major difference occurs with the Proximity and Projective categories. To answer **RQ2**, we can observe that the 'Proximity' category ranked similarly for the VLMs and higher than MGA-Net. On the other hand, 'Projective' has a higher rank for MGA-Net compared to all VLMs. We can see that MGA-Net prefers geometric spatial relations like left of, on top of, etc., as it explicitly encodes the relative locations of bounding boxes, which helps represent such relations. On the other hand, the VLMs have a better ranking for ambiguous relation categories that do not specify a clear distance or geometric direction (e.g., by, close to). This is because the vision backbones of the VLMs utilize the entire image and help capture relations between a region in the image and its surrounding regions, unlike MGA-Net, which only receives the detected bounding boxes as input.

Table 7 shows the performance of all models and the OWL-ViT baseline for expressions having different numbers of spatial relations. We observe that VLMs perform considerably better for expressions with 0/1 spatial relations compared to expressions with 2/3 spatial relations. This proves that VLMs find it comparatively difficult to ground multiple spatial relations. Among the VLMs, GDINO has the least drop in accuracy as no. of relations increase, while DeepSeek has the maximum drop. However, MGA-Net utilizes its compositional learning architecture to handle multi-step reasoning, resulting in a similar performance for all categories.

An interesting observation is that the performance of the baseline considerably drops for the 'Two' and 'Three' categories, even though the spatial relations aren't being passed as input to the baseline. The reason might be that 41.4% of these images have multiple instances of objects, the impact of which is explained in the next section.

Now, to answer **RQ3**, we observe in Table 7 that for all models except GDINO and LLaVA, the performance is better for expressions with one spatial relation than no spatial relations. Table 5 further shows that among the seven categories of single spatial relations, GDINO and LLaVa perform better for 4-5 of them compared to expressions with no relations. Thus, we can conclude that in a setup involving visual and linguistic ambiguity (such as ours), spatial relations along with visual attributes often aid the models in grounding the expressions, compared to the attributes alone. This is also reinforced by the results of the baseline. From Table 7, we can observe that while the baseline gives the second-best performance for expressions with no spatial relations, it drops to the third place for expressions with one spatial relation, with a 7.3% reduction in performance. This is because the baseline doesn't have access to the spatial relations.

Finally, Table 7 helps us answer **RQ4** as it shows the effect of increasing spatial relations on the per-

No. of relations	MGA-Net	GDINO	LLaVA	DeepSeek	OWL-ViT	Qwen
None	62.39 ± 0.59	73.88 ± 0	42.89 ± 0.17	41.10 ± 0	67.80 ± 0	66.14 ± 0
One	64.85 ± 0.13	73.94 ± 0	40.33 ± 0.01	42.53 ± 0	60.50 ± 0	71.40 ± 0
Two	61.96 ± 0.07	69.00 ± 0	31.05 ± 0.06	23.50 ± 0	52.39 ± 0	64.96 ± 0
Three	65.18 ± 2.1	67.78 ± 0	30.19 ± 0.26	11.11 ± 0	55.00 ± 0	62.78 ± 0

Table 7: Spatial relation frequency results and ranking

formance of MGA-Net versus the VLMs (as discussed before).

5.4 Impact of Visual Complexity

Model	Acc Single	Acc Multi
MGA-Net	64.91 ± 0.15	59.61 ± 0.04
G-DINO	72.54 ± 0.01	68.94 ± 0.01
LLaVA	37.69 ± 0.01	30.43 ± 0.1
DeepSeek-VL2	32.53 ± 0	27.46 ± 0
Qwen2.5-VL	68.47 ± 0	63.76 ± 0
OWL-ViT	59.71 ± 0	51.30 ± 0

Table 8: Results for accuracy in different visual complexity settings. All accuracies are in (%).

Out of 12586 test instances, we found that 4730 instances have images with multiple instances of objects mentioned in the referring expressions. Table 8 shows the accuracies of all models and the OWL-ViT baseline for images with a single instance ('Acc Single' column) and multiple instances ('Acc Multi' column). The models perform better for the single instance images by 5.7% on average compared to the multi-instance images. The 8.4% performance drop of the baseline for multi-instance images proves that the images are indeed complex and require more than just the label as the input for grounding the right object. Excluding the baseline, LLaVa has the steepest performance drop, showing that grounded pre-training also plays a crucial role in helping the models ground the right object instance in multi-instance images.

5.5 Impact of Negation

	2 Negations	1 Negation
Negation failure	92.86	93.75
DeepSeek	88.1	87.5
GDINO	78.57	62.5
LLaVA	35.71	25
MGA-Net	52.38	50
Qwen2.5-VL	50	62.5

Table 9: Results for negation analysis. All values are in (%) w.r.t the total failure instances (42 for 2 Negations and 16 for 1 Negation).

We obtained 16 expressions with 1 'not' and 42 expressions with 2 'not's for which all models gave incorrect predictions (as shown in Table 9). The 'Negation failure' row gives the percentage of instances for which at least 1 model gives an incorrect prediction due to failing to recognize negations and not due to conflation of other errors. We can observe that DeepSeek has the highest percentage of failure instances, while GDINO records the second most failures. However, while DeepSeek has the worst comprehension accuracy for the REC task, GDINO has the highest accuracy (refer to Table 4).

A possible reason for VLMs like LLaVA and Qwen2.5-VL performing better than GDINO in recognizing negations is due to their superior language backbones (Vicuna (novita.ai, 2024) and Qwen-2.5 LLM, respectively) that have better language understanding (including negations) compared to GDINO's CLIP text encoder. MGA-Net outperforms GDINO since its training involves expressions with negations, increasing its ability to comprehend negations during testing. Hence, to answer **RQ5**, we observe that while all REC models face issues with recognizing negations, certain model characteristics and training paradigms might reduce the failure cases.

Models	Negations	Precision	Recall
MGA-Net	1	53.60	70.8
MGA-Net	2	41.38	51
LLaVA	1	64.54	47.23
LLaVA	2	60.35	41

Table 10: Negation Precision (%) and Recall (%): MGA-Net vs. LLaVA

Another interesting observation was for the outputs of MGA-Net and LLaVA models when they are close to the target object. From Table 10, we can see that while LLaVA has a better precision in such cases, MGA-Net has a better recall.

6 Qualitative Analysis

In this section, we provide a qualitative analysis of certain issues faced by the models in handling

referring expressions.

6.1 Directional Relations

The expressions pertaining to Figures 1a and 1b consist of the same spatial relation ('around'). In the first figure, the wrapping of the napkin around the hot dog only makes the napkin partially visible. But in the second figure, the white box around the mirror is almost entirely visible. This shows how the interpretation of 'around' is highly dependent on the configuration of the involved objects. For the first image, LLaVA fails to precisely localize the object, while MGA-Net only returns a part of the napkin that is visible. In the second image, both models fail to localize the object. DeepSeek fails to output a bounding box for the first image but gets it right for the second.

6.2 Projective and Proximity Relations

Figure 1c shows an example of a Projective relation ('to the left'). MGA-Net succeeds in returning the correct part of the table that is to the left of the phone. While GDINO, DeepSeek, and Qwen simply return the entire table, LLaVA identifies the wrong part. This shows the ability of MGA-Net to comprehend projective relations better, particularly when the target object is not apparent. An example of Proximity relations is in Figure 1d, where LLaVA, GDINO, and DeepSeek return the shore that is 'near' the murky water, but MGA-Net fails to do so. Interestingly, Qwen only returns that part of the shore which isn't occluded by the elephants.

6.3 Multiple Spatial Relations

For 'Two-and' category expressions, the models sometimes only satisfy one of the spatial clauses. This often happens if multiple objects of the same class are in the image. For example, in Figure 1e's prediction for all models (except DeepSeek, which doesn't return any output), the output baseball player is to the left of the black helmet but is not to the right of the home plate.

Similarly, for 'Two-chained' category expressions, the models sometimes do not consider the entire expression. For example, in Figure 1f, MGA-Net, LLaVA, DeepSeek, and Qwen return the 'log that is behind the large bear', and GDINO returns the bear itself. None of the models consider the 'large branch' part of the expression, which should have been the output.

Finally, for 'Two-or' category expressions, the models might consider only one spatial clause. Consequently, it returns an object satisfying that clause but not the additional attributes mentioned in the expression. For example, in Figure 1g, the models return the monitor that is to the 'left of the keyboard', but it does not satisfy the color attribute.

6.4 Negation

Figures 1h and 1i show two cases where all models fail to recognize negation. In 1h, we can observe that while MGA-Net is wrong, LLaVA is close to the ground truth but partially covers the target object (high precision, low recall). In 1i, while LLaVA, GDINO and Qwen are wrong, MGA-Net is closest to the ground truth but covers an excess area (low precision, high recall).

7 Conclusion

Spatial reasoning is an integral aspect of cognitive reasoning and embodied AI tasks. However, recent studies have shown that state-of-the-art VLMs often fail to accurately comprehend spatial relations. To analyze the limitations of these models, we evaluate their spatial understanding using the referring expression comprehension task. We picked multiple models, including Vision-language models (LLaVA, GDINO, DeepSeek, Qwen) and task-specific models (MGA-Net). We observed that the VLMs that are trained in the wild with visual and textual data perform worse in grounding. The models perform the worst in grounding Directional relations on average. However, the VLMs do better in vague relations such as proximity, while the task-specific models are better in geometrically well-defined relations such as left and right. While using spatial relations increases the grounding accuracy, using multiple relations makes the reasoning more challenging for all models, with a higher impact on VLMs. However, MGA-Net maintains its performance for complex spatial expressions due to its compositional learning architecture. In the presence of visual complexity, the performance of all models drops, but DeepSeek and LLaVa's performances are affected the most due to a lack of grounded pre-training with complex expressions. Finally, both VLMs and task-specific models have failure cases when grounding expressions that include negation. These findings shed light on the gaps for future work on Vision-language models.

8 Future Directions

While we consider a comprehensive set of spatial relations for spatial analysis, these relations are atomic and do not include metric properties such as distances. We plan to extend our analysis to include metric relations in the future. Our results showed that increasing the number of parameters in VLMs can improve performance on expressions with simple spatial relations, architectural changes are needed to handle novel, complex compositions effectively. Unlike VLMs, MGA-Net maintains consistent performance across spatial complexities by using a soft attention module that decomposes expressions into semantic components for compositional reasoning. This suggests expression decomposition can enhance VLM generalization. Alternative strategies (Sinha et al., 2024) could be using multi-modal transformer models (Sikarwar et al., 2022; Qiu et al., 2021) and techniques such as weight sharing across transformer layers or 'Pushdown layers' with recursive language understanding (Murty et al., 2023). Another promising direction is Neuro-symbolic processing (Kamali et al., 2024; Hsu et al., 2024; Kamali and Kordjamshidi, 2025), which involves generating symbolic programs from expressions using LLMs and conducting explicit symbolic compositions before grounding into visual modality. Another challenge is the VLMs' inability to comprehend negations. Our experiments with the VLMs and MGA-Net suggest that augmenting the training/instruction tuning with synthetically generated negated expressions can alleviate this issue.

Acknowledgment

We would like to thank the anonymous reviewers for their insightful and thorough feedback. This presented research is partially supported by the Office of Naval Research (ONR) grant N00014-23-1-2417. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of Office of Naval Research.

Limitation

This paper provides an analysis of the shortcomings of the vision and language models when it comes to fine-grained spatial reasoning and grounding. Our analysis covers a variety of vision and language models including closed and open ones. However, the number of language models that we cover is by no means exhaustive. Spatial reasoning is important for many downstream applications however, we chose referring expressions as a platform that can demonstrate the challenges in both language and vision sides. While spatial understanding becomes a very important skill for embodied AI, in this work we do not consider the interaction with the environment and we do not consider the change of perspective. Our study can serve as a complement studies in this area that can provide insight into the difficulties of spatial language understanding and grounding language into visual perception. Our study was constrained by the cost of proprietary LLMs and the computational resources for open source ones.

References

Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian Reid, Stephen Gould, and Anton Van Den Hengel. 2018. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3674–3683.

Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Shiori Sagawa, and 1 others. 2023. OpenFlamingo: An Open-Source Framework for Training Large Autoregressive Vision-Language Models. *arXiv preprint arXiv:2308.01390*.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, and 1 others. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Jierun Chen, Fangyun Wei, Jinjing Zhao, Sizhe Song, Bohuai Wu, Zhuoxuan Peng, S-H Gary Chan, and Hongyang Zhang. 2024. Revisiting referring expression comprehension evaluation in the era of large multimodal models. *arXiv preprint arXiv:2406.16866*.

Zhenfang Chen, Peng Wang, Lin Ma, Kwan-Yee K Wong, and Qi Wu. 2020. Cops-Ref: A new Dataset and Task on Compositional Referring Expression Comprehension. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10086–10095.

Wenliang Dai, Junnan Li, D Li, AMH Tiong, J Zhao, W Wang, B Li, P Fung, and S Hoi. 2023. InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning. *arXiv preprint arXiv:2305.06500*, 2.

Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A Smith, Wei-Chiu Ma, and Ranjay Krishna. 2025. BLINK: Multimodal Large Language Models Can See but Not

Perceive. In *European Conference on Computer Vision*, pages 148–166. Springer.

Tejas Gokhale, Hamid Palangi, Besmira Nushi, Vibhav Vineet, Eric Horvitz, Ece Kamar, Chitta Baral, and Yezhou Yang. 2022. Benchmarking Spatial Relationships in Text-to-Image Generation. *arXiv preprint arXiv:2212.10015*.

Joy Hsu, Jiayuan Mao, Josh Tenenbaum, and Jiajun Wu. 2024. What's Left? Concept Grounding with Logic-Enhanced Foundation Models. *Advances in Neural Information Processing Systems*, 36.

Md Mofijul Islam, Alexi Gladstone, Riashat Islam, and Tariq Iqbal. 2023. EQA-MX: Embodied Question Answering using Multimodal Expression. In *The Twelfth International Conference on Learning Representations*.

Md Mofijul Islam, Reza Mirzaiee, Alexi Gladstone, Haley Green, and Tariq Iqbal. 2022. CAESAR: An Embodied Simulator for Generating Multimodal Referring Expression Datasets. *Advances in Neural Information Processing Systems*, 35:21001–21015.

Danial Kamali, Elham J Barezi, and Parisa Kordjamshidi. 2024. NeSyCoCo: A Neuro-Symbolic Concept Composer for Compositional Generalization. *arXiv preprint arXiv:2412.15588*.

Danial Kamali and Parisa Kordjamshidi. 2025. Neptune: A neuro-pythonic framework for tunable compositional reasoning on vision-language. *Preprint*, arXiv:2509.25757.

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

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, and 1 others. 2017. Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations. *International journal of computer vision*, 123:32–73.

Alexander Kuhnle, Huiyuan Xie, and Ann Copestake. 2018. How clever is the FiLM model, and how clever can it be? In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pages 0–0.

Martha Lewis, Nihal V Nayak, Peilin Yu, Qinan Yu, Jack Merullo, Stephen H Bach, and Ellie Pavlick. 2022. Does CLIP Bind Concepts? Probing Compositionality in Large Image Models. *arXiv preprint arXiv:2212.10537*.

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, Qingyang Wu, and Yong Jae Lee. 2024. Visual Instruction Tuning. *Advances in neural information processing systems*, 36.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and 1 others. 2023b. 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, and 1 others. 2023c. MMBench: Is Your Multi-modal Model an All-around Player? *arXiv preprint arXiv:2307.06281*.

Cristiane Kutianski Marchi Fagundes, Kristin Stock, and Luciene Stamato Delazari. 2021. A cross-linguistic study of spatial location descriptions in New Zealand English and Brazilian Portuguese natural language. *Transactions in GIS*, 25(6):3159–3187.

Matthias Minderer, Alexey Gritsenko, Austin Stone, Maxim Neumann, Dirk Weissenborn, Alexey Dosovitskiy, Aravindh Mahendran, Anurag Arnab, Mostafa Dehghani, Zhuoran Shen, and 1 others. 2022. Simple Open-Vocabulary Object Detection with Vision Transformers. In *European Conference on Computer Vision*, pages 728–755. Springer.

Shikhar Murty, Pratyusha Sharma, Jacob Andreas, and Christopher D Manning. 2023. Pushdown Layers: Encoding Recursive Structure in Transformer Language Models. *arXiv preprint arXiv:2310.19089*.

novita.ai. 2024. Vicuna: an Open-Source Large Language Model for Chatbots. Available at: https://blogs.novita.ai/vicuna-an-open-source-large-language-model-for-chatbots/. Published: 2024-04-18. Accessed: 2024-07-26.

Yanyuan Qiao, Chaorui Deng, and Qi Wu. 2020. Referring expression comprehension: A survey of methods and datasets. *IEEE Transactions on Multimedia*, 23:4426–4440.

Linlu Qiu, Hexiang Hu, Bowen Zhang, Peter Shaw, and Fei Sha. 2021. Systematic Generalization on gSCAN: What is Nearly Solved and What is Next? *arXiv preprint arXiv:2109.12243*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

J Redmon. 2016. You Only Look Once: Unified, Real-Time Object Detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2016. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE transactions on pattern analysis and machine intelligence*, 39(6):1137–1149.

Philipp J Rösch and Jindřich Libovický. 2023. Probing the Role of Positional Information in Vision-Language Models. *arXiv preprint arXiv:2305.10046*.

Ankur Sikarwar, Arkil Patel, and Navin Goyal. 2022. When Can Transformers Ground and Compose: Insights from Compositional Generalization Benchmarks. *arXiv preprint arXiv:2210.12786*.

Sania Sinha, Tanawan Premsri, and Parisa Kordjamshidi. 2024. A Survey on Compositional Learning of AI Models: Theoretical and Experimental Practices. *arXiv preprint arXiv:2406.08787*.

Sanjay Subramanian, William Merrill, Trevor Darrell, Matt Gardner, Sameer Singh, and Anna Rohrbach. 2022. ReCLIP: A Strong Zero-Shot Baseline for Referring Expression Comprehension. *arXiv preprint arXiv:2204.05991*.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, and 1 others. 2023. Gemini: A Family of Highly Capable Multimodal Models. *arXiv preprint arXiv:2312.11805*.

Yunjie Tian, Tianren Ma, Lingxi Xie, Jihao Qiu, Xi Tang, Yuan Zhang, Jianbin Jiao, Qi Tian, and Qixiang Ye. 2024. ChatterBox: Multi-round Multimodal Referring and Grounding. *arXiv preprint arXiv:2401.13307*.

Jiayu Wang, Yifei Ming, Zhenmei Shi, Vibhav Vineet, Xin Wang, Yixuan Li, and Neel Joshi. 2024. Is A Picture Worth A Thousand Words? Delving Into Spatial Reasoning for Vision Language Models. *arXiv preprint arXiv:2406.14852*.

Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang, and 1 others. 2024. Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding. *arXiv preprint arXiv:2412.10302*.

Licheng Yu, Zhe Lin, Xiaohui Shen, Jimei Yang, Xin Lu, Mohit Bansal, and Tamara L Berg. 2018. MAttNet: Modular Attention Network for Referring Expression Comprehension. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1307–1315.

Wentao Yuan, Jiafei Duan, Valts Blukis, Wilbert Pumacay, Ranjay Krishna, Adithyavairavan Murali, Arsalan Mousavian, and Dieter Fox. 2024. Robopoint: A vision-language model for spatial affordance prediction for robotics. *arXiv preprint arXiv:2406.10721*.

Yihan Zheng, Zhiquan Wen, Mingkui Tan, Runhao Zeng, Qi Chen, Yaowei Wang, and Qi Wu. 2020. Modular Graph Attention Network for Complex Visual Relational Reasoning. In *Proceedings of the Asian Conference on Computer Vision*.

A Description of spatial categories

For our analysis, we utilize the spatial categories introduced by (Marchi Fagundes et al., 2021) and replace the 'Cardinal Direction' category with 'Absolute'. The descriptions and examples for the chosen categories are as follows:

1. **Absolute**: Consists of relations that describe the location of an object in an absolute manner and not in relation to another object.
   E.g.: man on the right that is standing and wearing gray pant

2. **Adjacency**: Consists of relations that describe the close, side-by-side positioning of two objects. They may or may not imply a particular direction.
   E.g.: The large poster that is leaning against the wall

3. **Directional**: Consists of dynamic action verbs / directional relations. They describe the movement or change in position of an object relative to other objects in the image. The interpretation of these relations heavily relies on the configuration of the involved objects and/or the dynamic spatial relationship between them.
   E.g.: The gray car that is driving down the road

4. **Orientation**: Consists of relations which describe the orientation of an object w.r.t another object.
   E.g.: The sitting dog that is facing the window that is to the right of the mirror

5. **Projective**: Consists of relations that indicate the concrete spatial relationship between two objects, i.e., these relations can be quantified in terms of the coordinates of the two objects.
   E.g.: The black oven that is above the drawer

6. **Proximity**: Consists of relations that indicate that two objects are near each other without giving a specific directional relationship.
   E.g.: The blue chair that is close to the white monitor

7. **Topological**: Consists of relations that indicate the broader arrangement or the containment of an object w.r.t another object
E.g.: The silver train that is at the colorful station

8. **Unallocated**: Consists of relations that cannot be allocated to any of the above categories.

B Kendall Tau Independence Test

To compare the models' performances across the categories, we employ a statistical test known as the Kendall Tau Independence Test. It evaluates the degree of similarity between two sets of ranks given to the same set of objects. We calculate the Kendall rank coefficient (τ), which yields the correlation between two ranked lists. Given τ value, we calculate the z statistic, which follows standard normal distribution, as:

$$z = 3 * \tau * \sqrt{n(n-1)}/\sqrt{2(2n+5)}. \quad (1)$$

Using the 2-tailed p-test at 0.05 level of significance, we test the following:

- **Null hypothesis**: There is no correlation between the two ranked lists.
- **Alternative hypothesis**: There is a correlation between the two ranked

C Additional Model Settings and Experiments

Model	Accuracy (%)
MGA-Net (Layer 3)	62.92 ± 0.11
MGA-Net (Layer 4)	61.30 ± 0.09
LLaVA – Short Prompt	34.96 ± 0.03
LLaVA – Long Prompt	33.79 ± 0.01

Table 11: Additional Experiment Results - Comprehension Accuracies

For LLaVa (Liu et al., 2024), we experimented with the following two prompts:

- **Short prompt**: (USER: <image>\n Give the bounding box for: "Referring Expression"\nASSISTANT:)
- **Long prompt**: (USER: <image>\n Provide the bounding box coordinates for the object described by the referring expression: "Referring Expression"\n ASSISTANT:)

Both prompts are similar in structure, but the latter prompt is more verbose.

For MGA-Net (Zheng et al., 2020), we experiment with the Resnet-101 backbone by experimenting with Layers 3 and 4 for the visual features.

The results of these experiments are in Table 11. For LLaVA, the shorter prompt gives a slightly better result. For MGA-Net, the features captured by Layer 3 give a better result. Hence, we consider these two variants in all our experiments.

MGA-Net Hyperparameters. Since MGA-Net is the only model we train in this paper, we provide the hyperparameters used. These hyperparameters are derived from (Zheng et al., 2020)'s work. The model is trained for 15 epochs, as validation performance begins to degrade beyond that point. Training uses the Adam optimizer with a learning rate of 1e-4, batch size of 30, and gradient clipping set to 0.3. The language encoder is a 2-layer Bi-LSTM with a hidden size of 512 and no dropout. Word embeddings are 512-dimensional. Visual features are extracted via ResNet101, and object features include both visual and normalized spatial information, processed through MLPs. The model employs Gated Graph Neural Networks (GGNNs) for multi-step relational reasoning, using 3 update steps.

Model parameters. Among the models evaluated, **LLaVA** is a general-purpose vision-language model that integrates an open-set vision encoder from CLIP with a language decoder. The model is trained end-to-end through general visual instruction tuning to align visual and linguistic modalities. The commonly used LLaVA variant based on LLaMA-7B consists of approximately 7 billion parameters. **Grounding DINO** (base variant) has around 188 million parameters, optimized for phrase grounding and open-vocabulary detection. **DeepSeek-VL2 Tiny** is a compact model with about 600 million parameters, balancing speed and performance for multimodal tasks. **Qwen-VL 2.5** builds on the Qwen2.5 architecture and has 7 billion parameters, suitable for complex visual-language understanding. **OWL-ViT**, depending on the ViT backbone, ranges from **87M (ViT-B)** to **300M+ (ViT-L)** parameters, designed for open-vocabulary object detection. Finally, **MGA-Net**, tailored for referring expression grounding on datasets like CLEVR-Ref+, is lightweight with only 15–20 million parameters, yet delivers competitive task-specific performance.

D Experiments with other VLMs

In our analysis, we also experimented with InstructBLIP (Dai et al., 2023) and OpenFlamingo (Awadalla et al., 2023) models for the REC task. These models are general-purpose VLMs with InstructBLIP working in the zero-shot model and OpenFlamingo in the few-shot mode. In this section, we discuss the prompts that we used for these two models and the outputs obtained for the prompts:

D.1 InstructBLIP

For InstructBLIP, we designed three prompts for the REC task. They are as follows:

1. Bounding Boxes: bounding box list; Referring Expression: Refexp; The index of the output bounding box is:
2. Bounding Boxes: bounding box list; Referring Expression: Refexp; The coordinates of the output bounding box are:
3. Provide the bounding box coordinates for: "Refexp"

In the prompts, the 'bounding box list' placeholder takes the coordinates of the detected bounding boxes in the image being passed as the input, along with indices for each bounding box, starting from '1'. But for the third prompt, the model has no access to pre-detected candidate bounding boxes in the image. While the expected output for the first prompt is the index of the correct bounding box, for the other 2 prompts it is the bounding box coordinates as the output.

The bounding box format is [x1, y1, x2, y2], where (x1, y1) is the bottom left corner and (x2, y2) is the top right corner of the box. The coordinate values are a fraction of the total length/width of the image according to the position of the coordinate.

Unfortunately, none of the prompts gave consistently correct outputs. The outputs were as follows:
Prompt 1: The outputs were mostly incorrect. Sometimes, the model also gave '0' as the output, even though it is not a valid index.
Prompt 2: The output did not return meaningful coordinates in most cases. But in the few instances that it did, they were mostly incorrect. Example outputs when the model could not return meaningful coordinates are:

- `{1: [0.16, 0.55], 2: [0.32, 0.47], 3: [0.55, 0.6], 4: [0.21, 0.06]}`
- `[0.9, 0.53, 0.93, 0.57, 0.0, 0.39]`

Prompt 3: The model could not understand the task, and it just paraphrased parts of the prompt instead of giving the coordinates as the output. Example prompts and outputs are:

- **Prompt:** Provide the bounding box coordinates for: "The large poster that is leaning against the wall"
  Output: what is the bounding box coordinates for the large poster that is leaning against the wall
- **Prompt:** Provide the bounding box coordinates for: "The young man that is leaning against the wall"
  Output: is standing in an elevator. the young man that is leaning against the wall is standing in an elevator

D.2 OpenFlamingo

We tested all the prompts designed for OpenFlamingo in both 2 and 3-shot settings.

Prompt 1:

- **Example output format:** <image>Bounding Boxes:bounding box list; Expression: Refexp; Correct Bounding Box:"ID"<|endofchunk|>
- **Query format:** <image>Bounding Boxes:bounding box list; Expression: Refexp; Correct Bounding Box:"

'bounding box list' placeholder takes the list of candidate bounding boxes in the image as input, in the same format as InstructBLIP (discussed in the previous section). The expected output is the index of the correct bounding box. However, we observed that irrespective of the query, the model gave the same output index for the same set of prompting examples.

Prompt 2:

- **Example output format:** <image>Expression: Refexp; Correct Bounding Box:[Bounding box coordinates]<|endofchunk|>
- **Query format:** <image>Expression: Refexp; Correct Bounding Box:[

‘bounding box list’ placeholder takes the same input as explained for Prompt 1. But instead of expecting the index, we expect the coordinates of the bounding box as the output. The format of the bounding box is the same as explained for InstructBLIP in the previous section. However, the model failed to give meaningful coordinates as output in most cases. When it did give meaningful coordinates, the outputs were mostly incorrect.