CVPR
#12303

Keep it SymPL: Symbolic Projective Layout for Allocentric Spatial Reasoning in Vision-Language Models

Anonymous CVPR submission

Paper ID 12303

Abstract

*Perspective-aware spatial reasoning involves understanding spatial relationships from specific viewpoints—either egocentric (observer-centered) or allocentric (object-centered). While vision–language models (VLMs) perform well in egocentric settings, their performance deteriorates when reasoning from allocentric viewpoints, where spatial relations must be inferred from the perspective of objects within the scene. In this study, we address this underexplored challenge by introducing **Sym**bolic **P**rojective **L**ayout (SymPL), a framework that reformulates allocentric reasoning into symbolic-layout forms that VLMs inherently handle well. By leveraging four key factors—projection, abstraction, bipartition, and localization—SymPL converts allocentric questions into structured symbolic-layout representations. Extensive experiments demonstrate that this reformulation substantially improves performance in both allocentric and egocentric tasks, enhances robustness under visual illusions and multi-view scenarios, and that each component contributes critically to these gains. These results show that SymPL provides an effective and principled approach for addressing complex perspective-aware spatial reasoning.*

1. Introduction

Vision–language models (VLMs) have rapidly evolved, bridging visual perception and language understanding through large-scale multimodal learning [1, 10, 34]. Early models excelled in perception-focused tasks such as visual question answering and image captioning [15, 18, 47]. These advances have transformed VLMs from perception-oriented systems into general reasoning engines capable of understanding spatial structures and interacting with the physical world. In particular, spatial reasoning, which involves interpreting and inferring object relationships in three-dimensional (3D) space, has become essential for embodied artificial intelligence (AI) systems, including manipulation [14, 26] and navigation [11, 42].

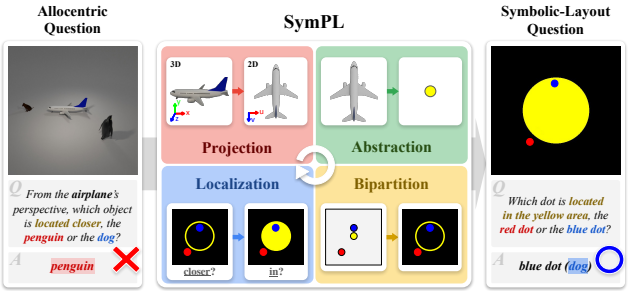

Figure 1. SymPL reformulates allocentric questions into symbolic-layout questions using four factors-*projection*, *abstraction*, *bipartition*, and *localization*-enabling significantly improved spatial reasoning under allocentric settings.

Despite recent progress, the spatial reasoning capability of VLMs remains limited. Prior studies [5–7, 9] have achieved partial success in egocentric reasoning, where relationships are interpreted from the observer's viewpoint, but performance drops sharply in allocentric settings, which require reasoning from the perspective of objects within the scene. This gap largely stems from the strong egocentric bias in existing training datasets [33, 50], causing models to underperform when confronted with viewpoint transformations. Such limitations hinder the use of VLMs in real-world tasks requiring multi-view or object-centered reasoning, such as autonomous driving [43] and robotic manipulation [36, 46].

Several strategies have been explored to address the limitations of allocentric spatial reasoning. A natural approach is to train models from scratch on allocentric datasets [22, 51], mirroring how egocentric spatial skills are typically acquired. Although such models can reach performance levels comparable to egocentric reasoning, the scarcity of allocentric data and the high computational cost make this approach impractical at scale. Fine-tuning pre-trained VLMs offers a more feasible alternative [36], but these models often generalize poorly beyond training contexts and suffer from catastrophic forgetting when new relations are introduced. Given these limitations, recent work has fo-

cused on improving allocentric reasoning without additional training, leveraging the inherent capabilities of large pre-trained VLMs. General reasoning aids such as chain-of-thought (CoT) [20] and visual prompting (VP) [23, 44] provide limited gains because they do not directly tackle viewpoint transformation. More recent approach [21] attempts to mitigate this challenge by transforming allocentric queries into egocentric ones—where VLMs tend to perform better—using auxiliary cues from foundation models [3, 19, 30, 41]. However, this conversion still underuses the intrinsic reasoning capacity of VLMs, leaving substantial room for improvement.

Building on these observations, we propose a new perspective: extending beyond viewpoint conversion, we transform the problem to leverage the factors that strongly influence VLM performance. To this end, we first analyze which factors correlate with higher answer accuracy in VLMs, and utilize these observations to reformulate allocentric reasoning in a way that maximizes their effectiveness. Through this analysis, we distill four fundamental factors that characterize effective spatial reasoning in VLMs:

- *Projection* : Spatial relations become easier for VLMs to process when projected onto a two-dimensional (2D) plane from an orthogonal viewpoint [25].
- *Abstraction* : Simplifying complex visual scenes into minimal, abstract symbols reduces irrelevant distractions, thereby enhancing reasoning robustness [2, 31].
- *Bipartition* : Spatial relationships are conveyed more intuitively when the reasoning space is minimally partitioned [16, 49].
- *Localization* : Asking whether an object lies within color-coded regions, rather than reasoning about direction or distance, improves inference accuracy [27, 37].

Building on these factors, we propose ***Sym**bolic **P**rojective **L**ayout* (SymPL), which reformulates allocentric questions into symbolic layouts that align with the reasoning patterns VLMs inherently handle well. As illustrated in Figure 1, SymPL estimates 3D information from the image and question, applies an orthogonal projection for relational comparison (*projection*), abstracts objects into minimal symbols (*abstraction*), partitions the space into two colored regions (*bipartition*), and converts the query into a position-estimation task (*localization*). These steps yield a concise *symbolic-layout question* that preserves essential spatial cues while filtering out redundant details, which can also be applied to egocentric spatial reasoning. Our experiments show that symbolic-layout questions substantially improve perspective-aware spatial reasoning accuracy and consistency. They preserve essential relational information while removing distracting cues, yielding clear gains in allocentric reasoning and also enhancing performance under egocentric, visual-illusion, and multi-view conditions. The contributions of this work are summarized as follows:

- We introduce SymPL, a method that optimizes complex allocentric spatial reasoning problems into forms where VLMs naturally excel.
- SymPL transforms allocentric questions into symbolic-layout questions using four key factors: *projection*, *abstraction*, *bipartition*, and *localization*.
- Experiments demonstrate that SymPL improves accuracy and ensures consistent, robust perspective-aware spatial reasoning across both allocentric and egocentric settings.

2. Related Works

2.1. Egocentric Spatial Reasoning

Egocentric spatial reasoning focuses on understanding spatial and geometric relationships from an observer-centered perspective [17, 35, 40]. Accurate reasoning of this kind relies on the alignment between visual relations in the image and their textual descriptions [8]. However, most VLMs do not enforce this alignment during training, yielding weak performance in such reasoning [1, 28]. To address this issue and enhance this capability, several studies have been proposed. SpatialVLM builds a large egocentric spatial reasoning dataset with 3D information and fine-tunes models on it [6]. SpatialRGPT injects object-level segmentation masks during training [9]. SpatialBot uses a progressive pipeline to integrate multi-level spatial cues [5]. SD-VLM learns from RGB and depth using metrically annotated 3D data [7]. Yet, these methods reason only from the egocentric views, and their performance degrades when the questions are from the allocentric views.

2.2. Allocentric Spatial Reasoning

Allocentric spatial reasoning refers to the ability to understand spatial relationships from the viewpoints of other objects in the scene rather than from the observer's perspective [2, 13]. This ability is essential for real-world applications that require understanding spatial relations based on surrounding objects, such as autonomous driving [43] and robotic interaction [36, 46]. Accordingly, several studies have been proposed to improve this reasoning capability. Some approaches address this problem by training models with allocentric datasets, yet they rely on costly data such as multi-view images or videos and learn only limited positional relations [22, 36]. As an alternative, APC reformulates the task into an egocentric form without additional training [21], leveraging auxiliary information extracted from foundation models [3, 19, 30, 41]. However, this conversion still falls short of leveraging the intrinsic reasoning capacity of VLMs.

2.3. General Reasoning Aids for Exploiting VLMs

Several reasoning aids have been proposed to enhance VLMs' reasoning capabilities. Chain-of-Thought (CoT)

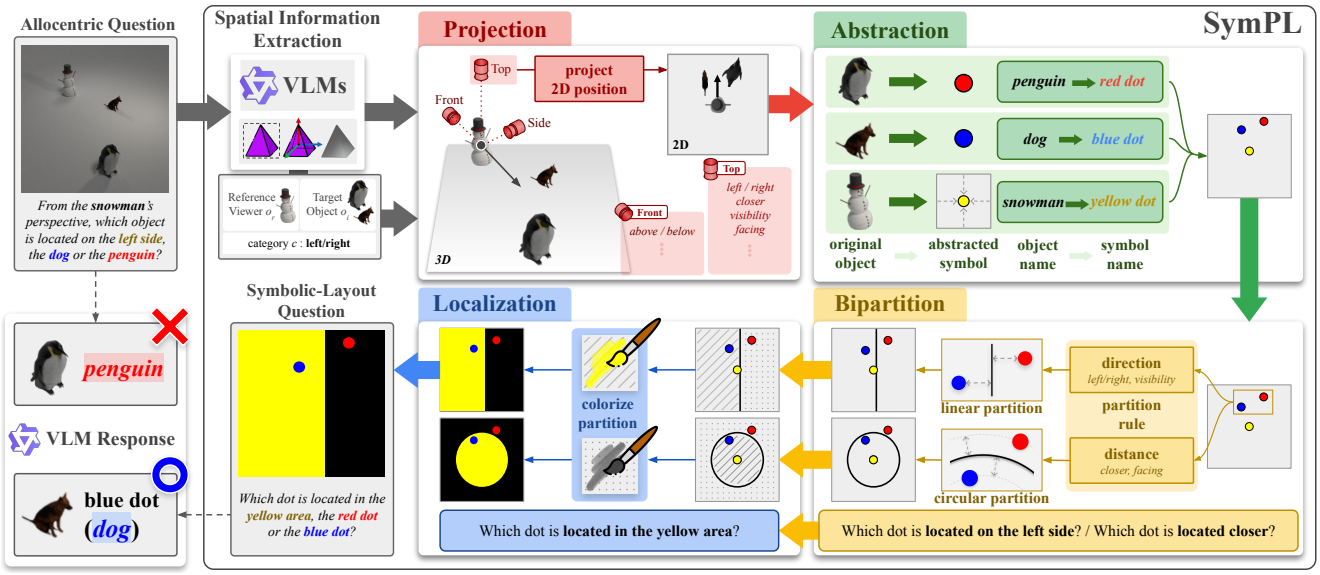

Figure 2. Overview of SymPL framework. SymPL reformulates an allocentric question into a symbolic-layout question through two stages: 1) Spatial Information Extraction and 2) Question Reformulation using four key factors — *projection*, *abstraction*, *bipartition*, and *localization*.

prompting guide models to perform complex multi-step reasoning through step-by-step examples, achieving strong performance in arithmetic and symbolic reasoning tasks [20]. Visual prompting is an annotation-based technique that helps VLMs focus on specific visual regions [4, 32, 38, 45, 48]. This approach encourages the model to focus more on the highlighted regions during reasoning. SoM prompting [44] introduces a method that applies semi-transparent segmentation masks to an image to distinguish semantic regions. SCAFFOLD prompting [23] marks regularly spaced dots on an image, intuitively visualizing object positions and clarifying spatial structure for the model. However, it remains unclear whether these approaches effectively enhance allocentric spatial reasoning.

3. Methodology

Our framework, SymPL, reformulates spatial reasoning problem from the allocentric view into an intuitive symbolic-layout question. As illustrated in Figure 2, SymPL performs reasoning through two stages: Spatial Information Extraction and Question Reformulation using four key factors. In the first stage, SymPL extracts 3D information for each object using pretrained models and a VLM. Next, SymPL integrates four key factors into the 3D information, namely *projection*, *abstraction*, *bipartition*, and *localization*, and generates a symbolic-layout question. This question is used as input to a VLM instead of the original question, to indirectly infer the answer to that question. The detailed descriptions of each stage are provided in the following sections.

3.1. Spatial Information Extraction

Given an allocentric question Q, consisting of an input image I and a prompt T, the framework first classifies the objects that constitute Q. The objects are composed of the *reference viewer* o_r, which serves as the basis of the perspective, and the *target objects* o_i, which are the reasoning targets. We perform two step reasoning processes to complete this classification. In the first reasoning step, the VLM extracts all object names mentioned in the prompt T as a list. This list includes both reference viewer and target objects. In the second reasoning step, the VLM also identifies the reference viewer from this list and constructs the final object set $O = \{o_r, o_i \mid i = 1, 2, \ldots, n\}$. By default, the reference viewer is explicitly specified in the prompt T. However, for egocentric questions, the 'camera' is designated as the reference viewer.

In the next step, we estimate the 3D coordinates of all objects. First, we detect the bounding boxes $B = \{b_r, b_i \mid i = 1, 2, \ldots, n\}$ for each object using GroundingDINO [30] and estimate the depth map D for the image I using Depth-Pro [3]. For each object o_j, where $j \in \{r, 1, \ldots, n\}$, the corresponding bounding box b_j is applied to the depth map D, the pixels in this region are unprojected to 3D, and the median of these points defines the object's 3D position $p_j = (x_j, y_j, z_j)$. During this process, we identified the area with the highest density of depth values inside the bounding box and used it to select inliers. We then removed outliers to reduce background, which minimized non-object regions in the mask. To handle scale differences between the x, y coordinates and the z values from the depth map, we apply

CVPR
#12303

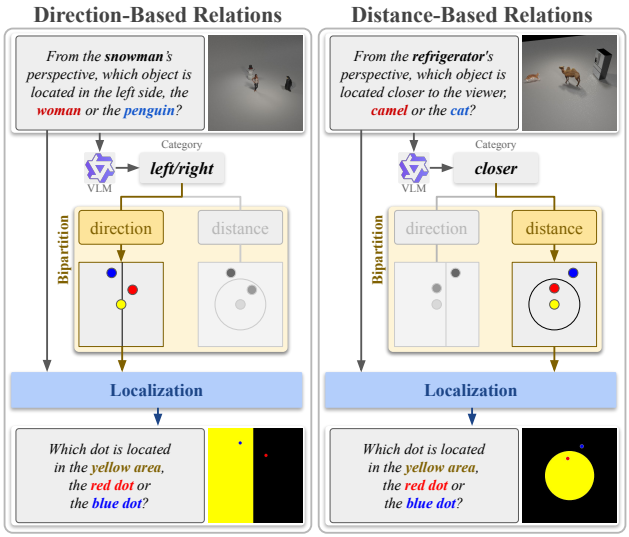

Figure 3. Partition rule based on spatial reasoning category. Directional comparisons adopt a linear partition, while distance comparisons employ a circular one.

a correction when the scale difference exceeds a predefined threshold to minimize distortion or bias in spatial relations.

In addition, the framework estimates the facing direction vector v_r of the reference viewer o_r in 3D space. First, the image I is cropped according to the reference viewer's bounding box b_r. The cropped image is then passed into the OrientAnything [41], which returns the facing direction vector v_r for the reference viewer. This result is combined with the object positions to form the 3D information set $U = \{v_r, p_r, p_i \mid i = 1, 2, \ldots, n\}$.

3.2. Question Reformulation using four key factors

This subsection describes how a symbolic-layout question is generated from the 3D information set U through four steps: *projection*, *abstraction*, *bipartition*, and *localization*. Each step is designed to apply the four key factors that characterize efficient spatial reasoning to allocentric questions. Before the steps, we first feed the prompt Q into the VLM to predict the spatial reasoning category c that needs to be inferred. The predicted category then serves as the guideline in each step for how the image should be reconstructed.

Projection Step. The projection step first chooses an external viewpoint centered on the reference viewer o_r and orthogonal to the plane where spatial relations are expressed. We choose the top view when reasoning category c is about relations on the plane on which the reference viewer stands, such as *left/right*, *closer*, *visibility*, and *facing*. We choose the front view for height related reasoning category such as *above/below*. Then, we project each object's 3D coordinate p_j to a 2D coordinate d_j in the chosen viewpoint. When projecting the 2D coordinate, we fix the reference viewer's facing direction v_r to the upward direction in the 2D plane so that the allocentric view maps to intuitive 2D spatial relations consistently. And we also fix the reference viewer's position p_r at the center of the 2D image, so that the target object is projected onto the image relative to that coordinate.

Abstraction Step. In this step, each original object is placed as an abstracted symbol based on the projected 2D coordinates d_j. We use this abstraction step to represent them in a way that stabilizes VLM recognition, because new-viewpoint reconstruction can fail to retain object shapes. During abstraction, objects are distinguished by unique colors so that the VLM recognizes symbols by color. We unify object shapes into featureless circles so that the VLM identifies symbols using color alone. The symbol for each object is then projected onto a 2D plane using d_j. Additionally, each object name is reconstructed as a color–shape combination that denotes an abstract symbol.

Bipartition Step. Next, the target object symbols are separated into a two-region layout within the abstracted image. As illustrated in Figure 3, the partition boundary is either linear or circular, and the partition form is determined by the type of spatial reasoning category to be inferred. If category c encodes directional comparisons, we use a linear partition so that different directions fall into distinct regions. For the *left/right* category, assuming the reference viewer's forward direction is projected upward in the image plane, we employ a vertical partition to separate left from right. For the *visibility* category, which determines whether the target object lies in front of or behind the reference viewer, we use a horizontal partition to separate front and back. Conversely, if category c requires comparing distances with respect to a specific location, we introduce a circular partition centered at that location to make distance differences visually distinguishable. For the *closer* category, the goal is to select the object closer to the reference viewer, so we partition the space using a circular boundary centered at the object of interest. For the *facing* category, which favors objects near the viewer's facing axis, we construct a circular partition centered at a point on that axis. Overall, this partition rule captures the key geometric cues and enable effective visualization of diverse spatial relations in the image space.

Localization Step. In the last step, questions about relative spatial relations are reformulated as a localization problem. First, we fill the bipartitioned regions with different colors. The colors used here are distinct from the target object symbol colors defined in the abstraction step. The painted colors serve to visually encode linguistic expressions of position. For example, if the reasoning category c is *left/right* and the abstracted image assigns yellow to the left region and black to the right region, then the positional expression 'left' can be represented by the visual cue 'yellow.' Accordingly, once the space is partitioned in this way, the positional relations mentioned in the input prompt T can likewise be converted into expressions about colors. Moreover, in this case, rela-

CVPR #12303

Group	Method	COMFORT#				3DSRBench		
		left/right	closer	visibility	facing	left/right	visibility	facing
Random	Random	48.75	48.67	47.27	52.33	50.72	50.00	47.69
General Purpose	LLaVA-NeXT [29]	47.33	51.58	51.80	52.42	37.39	51.89	58.53
	LLaVA-OneVision [24]	47.92	57.83	51.88	50.42	36.82	47.53	61.42
	Molmo [12]	46.58	46.17	53.36	<u>53.58</u>	39.97	50.29	59.39
	Qwen2.5-VL [1]	48.17	72.33	51.17	51.33	36.25	48.40	65.03
	Cambrian-1 [39]	41.17	80.42	50.70	41.33	40.83	53.63	67.05
	GPT-5 [34]	49.83	<u>84.25</u>	<u>54.22</u>	49.83	37.82	63.37	64.45
	Gemini-2.5-Flash [10]	38.33	77.83	52.34	51.58	38.40	64.10	**72.25**
General Reasoning Aids	Qwen2.5-VL + CoT [20]	43.25	70.75	50.39	44.67	33.52	51.74	63.44
	Qwen2.5-VL + SoM [44]	46.58	67.25	51.88	46.42	37.54	45.64	65.61
	Qwen2.5-VL + SCAFFOLD [23]	<u>52.17</u>	71.17	50.39	47.42	34.81	51.16	62.72
Egocentric Spatial Reasoning	SpatialVLM [6]	46.83	63.67	50.78	49.58	39.54	52.76	59.39
	SpatialRGPT [9]	43.08	70.25	53.75	47.75	36.53	49.56	62.57
	SpatialBot [5]	46.33	58.83	50.08	53.08	39.54	47.09	47.69
	SD-VLM [7]	45.83	45.17	48.91	52.25	49.71	46.95	48.84
Allocentric Spatial Reasoning	SAT [36]	35.00	48.75	34.92	39.50	44.56	34.45	25.43
	APC-Num [21]	47.83	52.50	34.14	36.92	<u>77.94</u>	56.10	58.24
	APC-Vis [21]	43.75	54.08	49.77	30.92	61.75	<u>71.37</u>	64.60
Ours	SymPL	**69.00**	**97.33**	**91.41**	**91.50**	**79.94**	**75.00**	<u>70.95</u>

Table 1. Quantitative results on allocentric questions. **Bold** indicates the best, while <u>underline</u> represents the second best results.

tive spatial questions like 'located on the left side' can be reduced to a localization problem such as 'located in' the color region that corresponds to that position. The resulting image–prompt pair becomes a symbolic-layout question Q^* that VLMs can effectively reason.

4. Experiments

In this section, we evaluated our framework, SymPL, across various spatial reasoning tasks and analyzed the utility of the symbolic layout question. To this end, we constructed five benchmark datasets and conducted experiments using diverse VLMs as baselines. In all experiments, we used the Qwen2.5-VL [1] for all reasoning in our framework.

4.1. Experiment Setup

4.1.1. Baseline VLMs

The reasoning performance of the SymPL framework was compared with that of various VLMs, and the models were categorized into four groups according to their roles and characteristics. The first group, referred to as General Purpose, includes both open-source and API-based VLMs [1, 10, 12, 24, 29, 34, 39]. These models represent general purpose VLMs with standard reasoning performances. Next, the General Reasoning Aids group includes a CoT method and two VP methods [20, 23, 44]. We incorporated these methods into the Qwen2.5-VL model. The Egocentric Spatial Reasoning group consists of four baselines, each fine-tuned for egocentric questions [5–7, 9]. The last group, referred to as Allocentric Spatial Reasoning, consists of three kinds of methods designed to solve allocentric questions [21, 36]. Additionally, we included a Random baseline that returns answers randomly to serve as a reference for comparison.

4.1.2. Datasets

To evaluate the performance of our model across diverse spatial reasoning tasks, we conducted experiments using five types of datasets. The detailed process for dataset construction is provided in the Supplementary Material.

COMFORT#. Focusing on allocentric spatial reasoning, this dataset is generated using a Blender-based synthetic dataset creation pipeline [50]. Each image is generated using an object set randomly selected from twelve types of object assets. The dataset consists of four categories: *left-/right*, *closer*, *visibility*, and *facing*.

3DSRBench. As a real-world benchmark for allocentric spatial reasoning, 3DSRBench is used in our experiments. We extracted only the categories that require considering allocentric view, namely *left/right*, *visibility*, and *facing*, for evaluation. The *visibility* category was adapted from the original dataset's *front/behind* category.

COCOSPATIAL. Serving as a real-world benchmark, COCOSPATIAL targets general egocentric spatial reasoning tasks. We constructed our benchmark using two categories from this dataset: *left/right* and *above/below*.

COMFORT VI. Built with the COMFORT# pipeline, COMFORT VI comprises spatial reasoning tasks under visual illusions. A visual illusion is a size-induced misperception of spatial relations. We constructed scenarios using col-

CVPR #12303

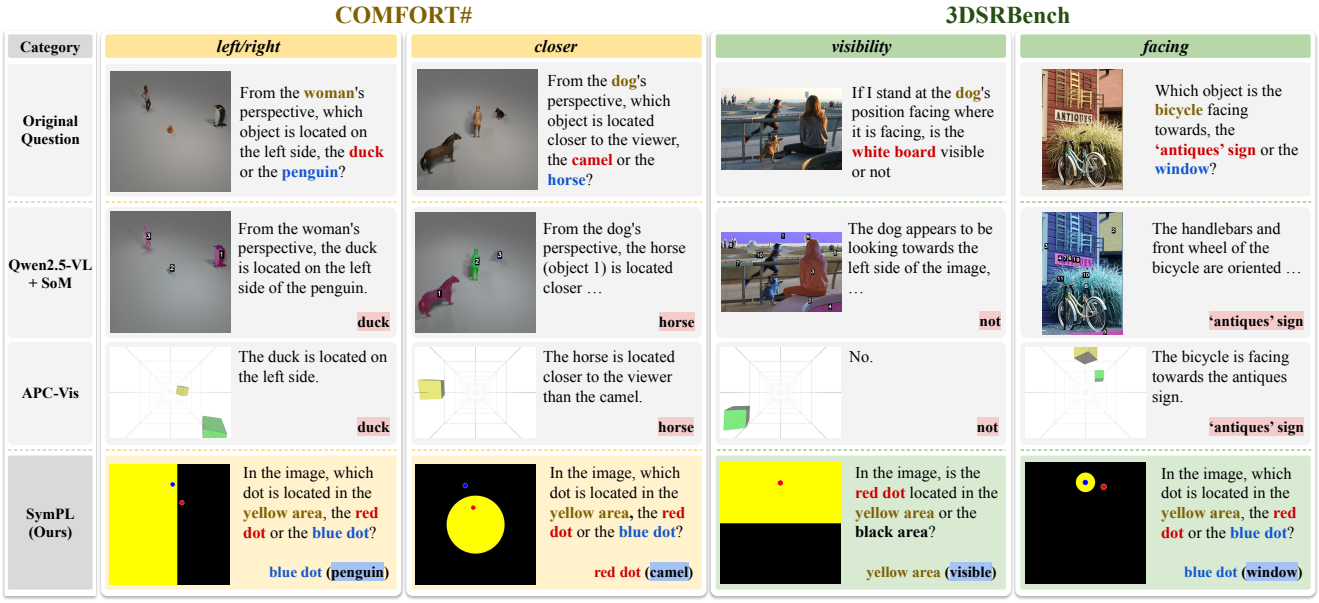

Figure 4. Allocentric spatial reasoning examples. Qwen2.5-VL+SoM and APC-Vis exhibited limited allocentric spatial reasoning performance across various categories. In contrast, our SymPL effectively handled allocentric questions by reformulating them into symbolic-layout questions.

ored balls of varying sizes and generated tasks per scenario, covered *left/right* for allocentric question and *front/behind* and *closer* for egocentric question.

COMFORT Multi. This dataset is designed to assess the model's consistency across viewpoints and was constructed using the COMFORT# pipeline. In this dataset, we captured 20 multi-view images per scene by moving the camera on a viewer-centered sphere with 72° azimuth and 15° polar steps, for 10 scenes per category. The dataset consists of four categories: *left/right*, *closer*, *visibility*, and *facing*.

4.2. Evaluation on Allocentric Questions

Evaluation on COMFORT#. As shown in Table 1, we evaluated whether the SymPL framework performs accurate reasoning across a range of allocentric questions. On COMFORT#, SymPL achieved *left/right* 69.00%, *closer* 97.33%, *visibility* 91.41%, and *facing* 91.50%, and outperformed all prior methods across categories. These results showed that the SymPL framework effectively performed allocentric spatial reasoning across multiple categories based on symbolic-layout questions. However, most baseline VLMs performed at or below the random baseline, and even GPT-5, which achieved the highest performance in the *closer* category as 84.25%, also performed worse in other categories. These findings indicated the limitations of VLMs' allocentric spatial reasoning and highlighted the importance of the proposed framework.

Evaluation on 3DSRBench. On 3DSRBench, SymPL achieved the highest accuracy in *left/right* at 79.94% and in *visibility* at 75.00%, and placed second in *facing* with 70.95%. Notably, many baselines tended to perform over 10% worse than the random baseline in the *left/right* category, showing that VLMs were biased toward an egocentric view. In the case of the APC series, the robustness of reasoning performance was low across various categories. Similarly, Gemini-2.5-Flash, which achieved the highest performance of 72.25% in the *facing* category, showed only 38.40% in the *left/right* category, underscoring weak

Method	COCOSPATIAL	
	left/right	above/below
Random	50.25	50.00
LLaVa-NeXT	<u>88.67</u>	76.00
LLaVa-OneVision	87.67	82.67
Molmo	81.17	72.17
Qwen2.5-VL	86.33	78.75
Cambrian-1	67.00	67.50
GPT-5	80.17	84.25
Gemini-2.5-Flash	88.58	<u>92.42</u>
Qwen2.5-VL + CoT	71.33	69.50
Qwen2.5-VL + SoM	79.58	74.17
Qwen2.5-VL + SCAFFOLD	84.33	81.92
SpatialVLM	78.92	71.00
SpatialRGPT	84.00	85.25
SpatialBot	84.50	83.50
SD-VLM	71.25	51.33
SAT	39.00	38.58
APC-Num	49.00	27.00
APC-Vis	49.92	54.17
SymPL (Ours)	**89.83**	**94.33**

Table 2. Quantitative results on egocentric questions. **Bold** indicates the best, while <u>underline</u> represents the second best results.

CVPR #12303

Method	COMFORT VI		
	Allocentric	Egocentric	
	left/right	front/behind	closer
Random	49.69	51.75	49.13
LLaVa-NeXT	58.19	58.75	89.13
LLaVa-OneVision	72.00	95.88	98.75
Molmo	44.56	93.88	74.50
Qwen2.5-VL	57.00	82.63	47.25
Cambrian-1	35.06	79.25	87.25
GPT-5	43.63	99.75	89.13
Gemini-2.5-Flash	54.63	55.50	44.00
Qwen2.5-VL + CoT	40.94	97.25	23.50
Qwen2.5-VL + SoM	46.56	44.25	46.13
Qwen2.5-VL + SCAFFOLD	58.06	85 00	68.50
SpatialVLM	44.00	92.13	80.75
SpatialRGPT	42.31	64.88	61.38
SpatialBot	52.25	78.25	87.00
SD-VLM	50.00	49.13	46.75
SAT	29.44	49.63	47.63
APC-Num	84.31	27.88	23.00
APC-Vis	76.75	36.25	42.25
SymPL (Ours)	**95.38**	**100.00**	**100.00**

Table 3. Quantitative results on perspective-aware reasoning under visual illusions. **Bold** indicates the best, while underline represents the second best results.

cross-category consistency. In contrast, SymPL's consistent gains across all categories demonstrated robustness to diverse spatial relations and validated the effectiveness of the symbolic-layout questions. Qualitative examples are provided in Figure 4.

4.3. Applying SymPL to Egocentric Questions

Next, the applicability of the SymPL framework for egocentric spatial reasoning was evaluated using COCOSPATIAL. The results are presented in Table 2. In this experiment, SymPL achieved 89.83% in *left/right* and 94.33% in *above/below*, which exceeded the best baselines. The results demonstrated that the symbolic-layout question yielded marked performance improvements even for egocentric questions. With most baselines having surpassed 70%, this showed that overall VLM performance was biased toward egocentric questions rather than allocentric ones. By contrast, APC-Num and APC-Vis showed a clear drop in accuracy, because the models are biased toward allocentric questions, misclassifying the camera viewpoint as allocentric rather than egocentric and generating incorrect scenes. These trends showed the effectiveness of the SymPL, which supports both allocentric and egocentric reasoning.

4.4. Assessing Perspective-Aware Reasoning under Visual Illusions

To assess spatial reasoning in more diverse scenarios, we conducted additional experiments using a dataset featuring induced visual illusions. In this setup, *left/right* is evaluated as an allocentric question, while *front/behind* and *closer* are egocentric. As shown in Table 3, SymPL achieved 100% on *front/behind* and *closer* and 95.38% on *left/right*, which demonstrated robust reasoning even under visual illusions. Additionally, among baseline VLMs, the APC series specialized in allocentric spatial reasoning, while the others were stronger in the egocentric setting. This suggested that most VLMs were biased toward one perspective or the other. By contrast, our framework again showed robust reasoning under visual illusions.

Method	left/right	closer	visibility	facing
Qwen2.5-VL	67.50	70.50	58.50	57.50
Qwen2.5-VL + CoT	57.50	66.00	60.50	50.50
Qwen2.5-VL + SoM	72.50	51.50	61.50	54.50
Qwen2.5-VL + SCAFFOLD	66.50	53.50	54.00	55.00
APC-Num	44.50	28.50	55.00	23.50
APC-Vis	53.00	31.50	61.50	16.50
SymPL (Ours)	**76.00**	**96.50**	**86.00**	**74.00**

Table 4. Quantitative results on viewpoint-aware consistency across multiple views. **Bold** indicates the best, while underline represents the second best results.

4.5. Evaluation of Viewpoint-Aware Consistency

An additional experiment was conducted to assess whether our approach yields consistent reasoning across images of the same scene captured from different viewpoints. As baselines, methods that exploit existing VLMs more effectively were considered, including CoT, SoM, SCAFFOLD, APC-Num, and APC-Vis. As shown in Table 4, SymPL achieved the highest success rate across all categories. These results indicated that our method supported robust allocentric reasoning that was invariant to the image-capture viewpoint.

4.6. Ablation Studies

4.6.1. Analysis of Each Key Factor

We analyzed the contribution of each of the four key factors to verify whether they have an effective impact on VLM reasoning. The experimental results are shown in Figure 5.

Projection. Viewpoint-dependent tendencies in spatial relation understanding were investigated by moving the camera from a front view to a top view while keeping the scene fixed. We built the scene using COMFORT# pipeline. We then analyzed the trend of reasoning performance changes according to viewpoint in the *above/below* category. As shown in Figure 5a, reasoning performance on *above/below* decreased as the camera approached a top view, indicating that selecting a viewpoint that matched the spatial relationships was crucial.

Abstraction. The effect of abstraction was examined in the *closer* category by comparing performance when the original images were annotated with segmentation masks against

CVPR
#12303

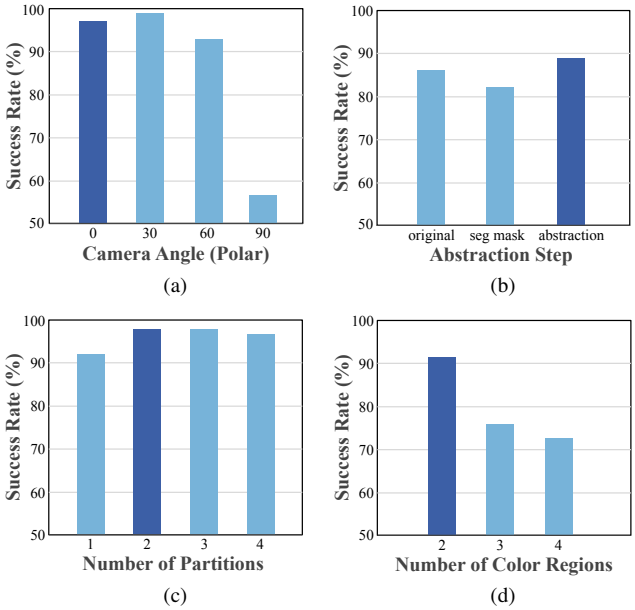

Figure 5. Ablation results of each key factor. (a) *projection*, (b) *abstraction*, (c) *bipartition*, (d) *localization*. The darker bar indicates the configuration used in SymPL.

performance when they were converted into abstract images. Figure 5b showed that expressing positional relationships in an abstracted form has a positive effect on reasoning compared to segmentation mask on the original image.

Bipartition. The impact of bipartition was evaluated in the *closer* category by measuring spatial reasoning performance as a function of the number of partitions. Figure 5c illustrated that the presence of partitions had a positive effect on reasoning, but the number of partitions made little difference in performance.

Localization. We analyzed reasoning performance as the number of regions distinguished by different colors increased using the *closer* category. In Figure 5d, performance dropped greatly as the number of regions to consider increased, confirming that partitioning a region into two parts and assigning different colors to them yields the notable gains.

4.6.2. Ablation on the Effectiveness of the Key Factor

An ablation study was conducted to examine how the performance on symbolic-layout questions changes as four key factors are added step by step. The experimental setup and results were summarized in Table 5. Starting from Setting 1, we sequentially incorporate *projection*, *abstraction*, *bipartition*, and *localization*, yielding the full symbolic-layout question at Setting 5. For evaluation, we randomly sampled 100 instances per category and tested five general purpose VLMs: Qwen2.5-VL, GPT-5, LLaVA-NeXT, LLaVA-OneVision, and Molmo. From the experimental results, Setting 1 exhibited the lowest performance in most

Setting	P	A	B	L	*left/right*	*closer*	*visibility*	*facing*
1					46.6	63.8	52.0	52.8
2	✓				89.2	64.8	51.2	52.0
3	✓	✓			96.4	81.0	90.8	100.0
4	✓	✓	✓		97.0	91.0	84.6	100.0
5 (Ours)	✓	✓	✓	✓	100.0	100.0	100.0	100.0

Table 5. Ablation results on the effectiveness of four key factors: **P**rojection, **A**bstraction, **B**ipartition, and **L**ocalization. Results show the average success rate of five general-purpose VLMs for each category: *left/right*, *closer*, *visibility*, and *facing*.

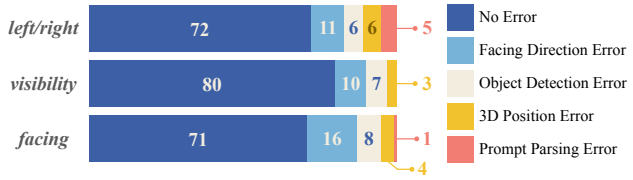

Figure 6. Error breakdown in SymPL's reasoning pipeline.

categories, and performance tended to improve as the stages progressed. Under Setting 5, all experiments uniformly achieved 100%. These results indicated that the four key factors acted synergistically and combining them improved reasoning performance.

4.7. Error Breakdown

We analyzed a manual error breakdown to assess failure cases in SymPL's reasoning pipeline, using inference results from 100 randomly sampled instances per 3DSRBench category. Figure 6 showed that the most frequent error across categories involved misestimating the reference viewer's facing-direction vector. Other common errors included incorrect object detection, inaccurate 3D coordinate estimation, and misidentification of object names specified in the prompt. Most failures appeared to arise from reliance on foundation models, and reasoning failures on the symbolic-layout question were not observed in this analysis.

5. Conclusion

In this paper, we introduce SymPL, a framework that optimizes complex spatial reasoning problem into forms where VLMs excel effectively. SymPL reformulates questions into intuitive symbolic-layout questions based on four key factors, enabling effective reasoning. Our experiments show that symbolic-layout questions significantly improve both allocentric and egocentric spatial reasoning. These results demonstrate that SymPL provides an effective and principled approach for addressing complex perspective-aware spatial reasoning.

References

[1] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 1, 2, 5

[2] Christopher Beckham, Martin Weiss, Florian Golemo, Sina Honari, Derek Nowrouzezahrai, and Christopher Pal. Visual question answering from another perspective: Clevr mental rotation tests. *Pattern Recognition*, 136:109209, 2023. 2

[3] Aleksei Bochkovskii, Amaël Delaunoy, Hugo Germain, Marcel Santos, Yichao Zhou, Stephan R. Richter, and Vladlen Koltun. Depth pro: Sharp monocular metric depth in less than a second. In *International Conference on Learning Representations*, 2025. 2, 3

[4] Mu Cai, Haotian Liu, Siva Karthik Mustikovela, Gregory P. Meyer, Yuning Chai, Dennis Park, and Yong Jae Lee. Making large multimodal models understand arbitrary visual prompts. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2024. 3

[5] Wenxiao Cai, Yaroslav Ponomarenko, Jianhao Yuan, Xiaoqi Li, Wankou Yang, Hao Dong, and Bo Zhao. Spatialbot: Precise spatial understanding with vision language models. *IEEE International Conference on Robotics and Automation (ICRA)*, 2024. 1, 2, 5

[6] Boyuan Chen, Zhuo Xu, Sean Kirmani, Brain Ichter, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14455–14465, 2024. 2, 5

[7] Pingyi Chen, Yujing Lou, Shen Cao, Jinhui Guo, Lubin Fan, Yue Wu, Lin Yang, Lizhuang Ma, and Jieping Ye. Sd-vlm: Spatial measuring and understanding with depth-encoded vision-language models. *Advances in Neural Information Processing Systems*, 2025. 1, 2, 5

[8] Shiqi Chen, Tongyao Zhu, Ruochen Zhou, Jinghan Zhang, Siyang Gao, Juan Carlos Niebles, Mor Geva, Junxian He, Jiajun Wu, and Manling Li. Why is spatial reasoning hard for vlms? an attention mechanism perspective on focus areas. In *International conference on machine learning*, 2025. 2

[9] An-Chieh Cheng, Hongxu Yin, Yang Fu, Qiushan Guo, Ruihan Yang, Jan Kautz, Xiaolong Wang, and Sifei Liu. Spatialrgpt: Grounded spatial reasoning in vision-language models. In *NeurIPS*, 2024. 1, 2, 5

[10] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025. 1, 5

[11] Jonathan Crespo, Jose Carlos Castillo, Oscar Martinez Mozos, and Ramon Barber. Semantic information for robot navigation: A survey. *Applied Sciences*, 10(2):497, 2020. 1

[12] Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, et al. Molmo and pixmo: Open weights and open data for state-of-the-art vision-language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 91–104, 2025. 5

[13] Paola Del Sette, Markus Bindemann, and Heather J Ferguson. Visual perspective-taking in complex natural scenes. *Quarterly Journal of Experimental Psychology*, 75(8):1541–1551, 2022. 2

[14] Jensen Gao, Bidipta Sarkar, Fei Xia, Ted Xiao, Jiajun Wu, Brian Ichter, Anirudha Majumdar, and Dorsa Sadigh. Physically grounded vision-language models for robotic manipulation. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 12462–12469. IEEE, 2024. 1

[15] Zonghao Guo, Ruyi Xu, Yuan Yao, Junbo Cui, Zanlin Ni, Chunjiang Ge, Tat-Seng Chua, Zhiyuan Liu, and Gao Huang. Llava-uhd: an lmm perceiving any aspect ratio and high-resolution images. In *European Conference on Computer Vision*, pages 390–406. Springer, 2024. 1

[16] Amirmohammad Izadi, Mohammadali Banayeeanzade, Fatemeh Askari, Ali Rahimiakbar, Mohammad Mahdi Vahedi, Hosein Hasani, and Mahdieh Soleymani Baghshah. Visual structures help visual reasoning: Addressing the binding problem in LVLMs. In *The Thirty-Ninth Conference on Neural Information Processing Systems (NeurIPS 2025)*, 2025. Poster. 2

[17] Amita Kamath, Jack Hessel, and Kai-Wei Chang. What's "up" with vision-language models? investigating their struggle with spatial reasoning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9161–9175, Singapore, 2023. Association for Computational Linguistics. 2

[18] Seil Kang, Jinyeong Kim, Junhyeok Kim, and Seong Jae Hwang. Your large vision-language model only needs a few attention heads for visual grounding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 9339–9350, 2025. 1

[19] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 2

[20] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213, 2022. 2, 3, 5

[21] Phillip Y Lee, Jihyeon Je, Chanho Park, Mikaela Angelina Uy, Leonidas Guibas, and Minhyuk Sung. Perspective-aware reasoning in vision-language models via mental imagery simulation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025. 2, 5

[22] Seonho Lee, Jiho Choi, Inha Kang, Jiwook Kim, Junsung Park, and Hyunjung Shim. 3d-aware vision-language models fine-tuning with geometric distillation. In *Findings of the*

Association for Computational Linguistics: EMNLP 2025, 2025. 1, 2

[23] Xuanyu Lei, Zonghan Yang, Xinrui Chen, Peng Li, and Yang Liu. Scaffolding coordinates to promote vision-language coordination in large multi-modal models. *Conference on Computational Linguistics*, 2024. 2, 3, 5

[24] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024. 5

[25] Chengzu Li, Caiqi Zhang, Han Zhou, Nigel Collier, Anna Korhonen, and Ivan Vulić. Topviewrs: Vision-language models as top-view spatial reasoners. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1786–1807, Miami, Florida, USA, 2024. Association for Computational Linguistics. 2

[26] Xiaoqi Li, Mingxu Zhang, Yiran Geng, Haoran Geng, Yuxing Long, Yan Shen, Renrui Zhang, Jiaming Liu, and Hao Dong. Manipllm: Embodied multimodal large language model for object-centric robotic manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18061–18070, 2024. 1

[27] Yijun Liang, Ming Li, Chenrui Fan, Ziyue Li, Dang Nguyen, Kwesi Adu Cobbina, Shweta Bhardwaj, Jiuhai Chen, Fuxiao Liu, and Tianyi Zhou. Colorbench: Can vlms see and understand the colorful world? a comprehensive benchmark for color perception, reasoning, and robustness. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. 2

[28] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023. 2

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

[30] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *European Conference on Computer Vision*, 2023. 2, 3

[31] Julius Mayer, Mohamad Ballout, Serwan Jassim, Farbod Nosrat Nezami, and Elia Bruni. iVISPAR — an interactive visual-spatial reasoning benchmark for VLMs. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 26745–26769, Suzhou, China, 2025. Association for Computational Linguistics. 2

[32] Soroush Nasiriany, Fei Xia, Wenhao Yu, Ted Xiao, Jacky Liang, Ishita Dasgupta, Annie Xie, Danny Driess, Ayzaan Wahid, Zhuo Xu, Quan Vuong, Tingnan Zhang, Tsang-Wei Edward Lee, Kuang-Huei Lee, Peng Xu, Sean Kirmani, Yuke Zhu, Andy Zeng, Karol Hausman, Nicolas Heess, Chelsea Finn, Sergey Levine, and Brian Ichter. Pivot: Iterative visual prompting elicits actionable knowledge for vlms. *International conference on machine learning*, 2024. 3

[33] Michael Ogezi and Freda Shi. SpaRE: Enhancing spatial reasoning in vision-language models with synthetic data. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7855–7875, Vienna, Austria, 2025. Association for Computational Linguistics. 1

[34] OpenAI. Gpt-5 system card. Technical report, OpenAI, 2025. Version as of Aug 2025. 1, 5

[35] Kanchana Ranasinghe, Satya Narayan Shukla, Omid Poursaeed, Michael S Ryoo, and Tsung-Yu Lin. Learning to localize objects improves spatial reasoning in visual-llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12977–12987, 2024. 2

[36] Arijit Ray, Jiafei Duan, Ellis Brown, Reuben Tan, Dina Bashkirova, Rose Hendrix, Kiana Ehsani, Aniruddha Kembhavi, Bryan A. Plummer, Ranjay Krishna, Kuo-Hao Zeng, and Kate Saenko. Sat: Dynamic spatial aptitude training for multimodal language models. In *Proceedings of the Conference on Language Modeling (COLM 2025)*, 2025. 1, 2, 5

[37] Fatemeh Shiri, Xiao-Yu Guo, Mona Golestan Far, Xin Yu, Gholamreza Haffari, and Yuan-Fang Li. An empirical analysis on spatial reasoning capabilities of large multimodal models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 21440–21455, Miami, Florida, USA, 2024. Association for Computational Linguistics. 2

[38] Aleksandar Shtedritski, Christian Rupprecht, and Andrea Vedaldi. What does clip know about a red circle? visual prompt engineering for vlms. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11987–11997, 2023. 3

[39] Peter Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Adithya Jairam Vedagiri IYER, Sai Charitha Akula, Shusheng Yang, Jihan Yang, Manoj Middepogu, Ziteng Wang, et al. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. *Advances in Neural Information Processing Systems*, 37:87310–87356, 2024. 5

[40] Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9568–9578, 2024. 2

[41] Zehan Wang, Ziang Zhang, Tianyu Pang, Chao Du, Hengshuang Zhao, and Zhou Zhao. Orient anything: Learning robust object orientation estimation from rendering 3d models. *International conference on machine learning*, 2024. 2, 4

[42] Steffen Werner, Bernd Krieg-Brückner, Hanspeter A Mallot, Karin Schweizer, and Christian Freksa. Spatial cognition: The role of landmark, route, and survey knowledge in human and robot navigation1. In *Informatik'97 Informatik als Innovationsmotor: 27. Jahrestagung der Gesellschaft für Informatik Aachen, 24.–26. September 1997*, pages 41–50. Springer, 1997. 1

[43] Dingkang Yang, Kun Yang, Yuzheng Wang, Jing Liu, Zhi Xu, Rongbin Yin, Peng Zhai, and Lihua Zhang. How2comm: Communication-efficient and collaboration-pragmatic multi-agent perception. *Advances in Neural Information Processing Systems*, 36:25151–25164, 2023. 1, 2

[44] Jianwei Yang, Hao Zhang, Feng Li, Xueyan Zou, Chunyuan Li, and Jianfeng Gao. Set-of-mark prompting unleashes extraordinary visual grounding in gpt-4v, 2023. 2, 3, 5

[45] Lingfeng Yang, Yueze Wang, Xiang Li, Xinlong Wang, and Jian Yang. Fine-grained visual prompting. *Advances in Neural Information Processing Systems*, 36:24993–25006, 2023. 3

[46] Zhutian Yang, Caelan Garrett, Dieter Fox, Tomás Lozano-Pérez, and Leslie Pack Kaelbling. Guiding long-horizon task and motion planning with vision language models. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pages 16847–16853. IEEE, 2025. 1, 2

[47] Qinghao Ye, Xianhan Zeng, Fu Li, Chunyuan Li, and Haoqi Fan. Painting with words: Elevating detailed image captioning with benchmark and alignment learning. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2025. Accepted at ICLR 2025. 1

[48] Runpeng Yu, Weihao Yu, and Xinchao Wang. Api: Attention prompting on image for large vision-language models. In *European Conference on Computer Vision*, 2024. 3

[49] Haoquan Zhang, Ronggang Huang, Yi Xie, and Huaidong Zhang. Mask4align: Aligned entity prompting with color masks for multi-entity localization problems. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13373–13383, 2024. 2

[50] Zheyuan Zhang, Fengyuan Hu, Jayjun Lee, Freda Shi, Parisa Kordjamshidi, Joyce Chai, and Ziqiao Ma. Do vision-language models represent space and how? evaluating spatial frame of reference under ambiguities. In *The Thirteenth International Conference on Learning Representations*, 2025. 1, 5

[51] Duo Zheng, Shijia Huang, Yanyang Li, and Liwei Wang. Learning from videos for 3d world: Enhancing mllms with 3d vision geometry priors. In *Advances in Neural Information Processing Systems*, 2025. 1