Perspective-Aware Reasoning in Vision-Language Models via Mental Imagery Simulation

Anonymous ICCV submission

Paper ID 17

Figure 1. We introduce **Abstract Perspective Change (APC)**, a framework that empowers VLMs to adopt arbitrary perspectives for spatial reasoning. As demonstrated by the examples above, APC significantly enhances VLM's ability to *imagine a scene from alternative viewpoints*, overcoming the inherent egocentric bias that constrains the spatial reasoning of existing VLMs to the camera's viewpoint.

Abstract

We present a framework for perspective-aware reasoning in vision-language models (VLMs) through mental imagery simulation. Perspective-taking—the ability to perceive an environment or situation from an alternative viewpoint—is a key benchmark for human-level visual understanding, essential for environmental interaction and collaboration with autonomous agents. Despite advancements in spatial reasoning within VLMs, recent research has shown that modern VLMs significantly lack perspective-aware reasoning capabilities and exhibit a strong bias toward egocentric interpretations. To bridge the gap between VLMs and human perception, we focus on the role of mental imagery, where humans perceive the world through abstracted representations that facilitate perspective shifts. Motivated by this, we propose a framework for perspective-aware reasoning, named Abstract Perspective Change (APC), that effectively leverages vision foundation models, such as object detection, segmentation, and orientation estimation, to construct scene abstractions and enable perspective changes. Our experiments on synthetic and real-image benchmarks, compared with various VLMs, demonstrate significant improvements in perspective-aware reasoning with our framework, further outperforming fine-tuned spatial reasoning models and novel-view-synthesis-based approaches.

1. Introduction

Vision-language models (VLMs) have made remarkable progress, positioning themselves as a crucial backbone for general-purpose physical AI agents. The growing research efforts to improve VLMs' spatial reasoning capabilities [8, 10, 45, 64] reflect this potential. Early VLMs were limited to basic tasks such as visual question answering (VQA) and image captioning [1, 34, 35]. However, recent advancements have enabled them to perform complex visual reasoning [2, 14, 23, 24, 40, 63] and extract spatial properties, including spatial relationships, relative sizes, and distances [31, 64]. Further techniques such as instruction-tuning and vision-centric adapters, have expanded their capabilities, allowing depth-aware [6, 8] and region-aware [10, 21] spatial reasoning.

Despite these advances, progress remains largely confined to *egocentric* spatial reasoning, and even the latest VLMs struggle with *allocentric* reasoning—answering questions from perspectives other than the camera's (Fig. 2). Allocentric reasoning is crucial for high-level planning, environmental interaction, and collaboration with autonomous agents [16, 70, 74]. Moreover, it serves as a key benchmark for human-level spatial understanding. However, as analyzed by Zhang et al. [78], most VLMs exhibit a strong bias toward an *egocentric* perspective. Even when explicitly prompted to adopt an *allocentric* viewpoint, VLMs often revert to egocentric interpretations [19, 37, 77, 78]. Recent efforts to enhance spatial reasoning remain focused on improving egocentric reasoning [8, 10, 45], leaving allocentric reasoning largely unaddressed.

To bridge the gap between VLMs and human perspective reasoning, we ask: *What cognitive process allows humans to effortlessly shift perspectives?* Unlike current VLMs, humans seamlessly form internal representations of the physical world, making perspective reasoning an intuitive and natural process. The mechanism of creating internal representations, known as *mental imagery* [17, 29, 48, 49, 57], plays a fundamental role in cognition, enabling us to simulate visual, spatial, and conceptual scenarios. This ability allows for abstraction beyond immediate perception, facilitating sophisticated spatial reasoning tasks such as mentally rotating objects, inferring occlusions, and envisioning alternative viewpoints [5, 12, 15].

A key aspect of mental imagery is that it is not simply the process of visualizing a clear image from different perspectives; rather, it involves forming an *abstract representation* of a scene that encodes essential spatial information and can be reinterpreted from a new perspective. From a computational standpoint, such an abstract representation is particularly advantageous, as equipping VLMs with the imaginative capability to generate novel views remains extremely challenging. In contrast, constructing an abstract representation requires significantly less computation and can be achieved procedurally.

Inspired by this, we introduce a novel framework for adapting perspectives in VLMs by simulating the mental imagery process and modifying the perspective in the given prompt. Our goal is to leverage the strengths of both VLMs and recent vision foundation models, such as object detection [7, 43, 81], image segmentation [27, 52], and orientation estimation [66]. The proposed framework takes an image and a perspective-based question as input and operates through three key stages. First, by simulating the mental imagery process, it builds an abstract representation of the scene in the input image. The VLM parses the prompt to identify objects in the image, while vision foundation models extract the center and orientation information of each object in 3D space. Second, the VLM analyzes the prompt to determine the reference object from whose perspective the question is asked, and transforms the abstraction to be aligned with that perspective. Finally, a new *prompt* is generated by reinterpreting the scene from the reference object's perspective. We explore different formats for rendering the scene abstraction: (1) a text-based representation, where objects are described using numerical 3D coordinates, and (2) an image-based representation, where objects are visualized as colored boxes corresponding to the original image. The newly generated prompt is fed to the VLM to obtain the final answer. Note that our framework is *not* designed for a specific type of allocentric question. We leverage the egocentric reasoning capabilities of VLMs by performing allocentric-to-egocentric prompt conversion through scene abstraction, removing the perspective-related barrier in the question while preserving its original intent in the new prompt.

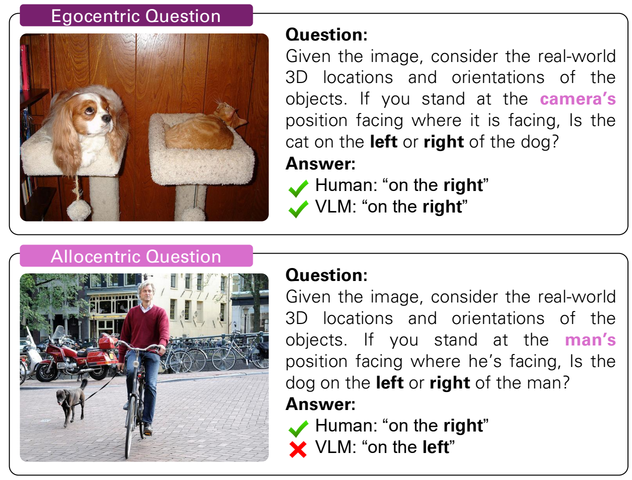

Figure 2. **Egocentric vs. Allocentric.** While VLMs perform well when questions are asked from an egocentric (*i.e.* camera's) perspective, they struggle when the same questions are posed from an allocentric perspective, showing a strong bias toward egocentric reasoning.

In our experiments on COMFORT++ [78] and 3DSR-Bench [46], our method achieves robust spatial reasoning across a variety of tasks and perspectives. In constrast, baseline VLMs and previous frameworks for spatial reasoning often struggle with even simple viewpoint shifts, reconfirming a notable bias toward the camera's perspective. These results highlight how our abstraction-based representation significantly enhances the spatial reasoning capabilities of VLMs beyond their default egocentric perspectives.

2. Related Work

2.1. Spatial Reasoning with VLMs

Building on the remarkable advancements of vision-language models (VLMs) [2, 14, 31, 40, 41], recent studies have adapted VLMs for real-world spatial reasoning. Numerous evaluations revealed that VLMs struggle on even elementary spatial-perception tasks [18, 50, 51, 62, 65] and higher-level spatial reasoning based on images or videos [26, 32, 39, 58, 60, 72]. SpatialVLM [8] tackles this issue with a data-synthesis pipeline that injects rich spatial cues, while Cambrian-1 [64] introduces an architecture purposed for improved spatial reasoning. Another line of work allows VLMs to utilize richer vision-centric data such as points, depth maps or segmentation masks through fine-tuning [6, 44, 60, 76] or employing auxiliary encoders [10]. Taking a different approach, other works exploit the planning and programming abilities of language models, building LLM/VLM-in-the-loop systems that call external vision modules as needed [20, 47, 61]. Notably, SpatialPIN [45] extracts dense visual priors from multiple vision foundation models [25, 42] and uses a VLM [24] to combine and interpret this information.

2.2. Visual Perspective-Taking

Visual perspective-taking (VPT) is the ability to imagine an alternate viewpoint, whether from another person's perspective or a different camera angle. This ability is essential for fundamental human skills such as navigation, spatial awareness, and social interaction [3, 11, 17, 55]. To be regarded as a general vision agent capable of human-like reasoning, a VLM should possess robust perspective-taking abilities. However, recent analyses reveal that current VLMs fail to shift to allocentric perspectives, showing a strong bias toward the egocentric viewpoint of a given image [19, 37, 46, 77, 78]. Zhang et al. [78] propose a synthetic evaluation protocol to assess whether VLMs can adopt different frames of reference (*i.e*. perspective). Likewise, 3DSRBench [46] includes real image-question pairs asked from an object's viewpoint, and finds that recent VLMs still demonstrate near chance level on perspective-related tasks. These findings suggest that while VLMs are rapidly improving in both complex visual reasoning [2, 14, 24, 40, 63] and basic spatial reasoning [6, 8, 10, 21, 31, 64], their abilities remain confined to the egocentric viewpoint, posing a significant barrier to human-like reasoning. Recently, SAT [53] proposed to improve VLMs' allocentric reasoning through instruction-tuning, yet it remains restricted to left/right relations with the need for annotations. In this work, we empower VLMs to reason from *arbitrary* perspectives, by reformulating any spatial reasoning task into their default egocentric viewpoint, resulting in a generalizable framework.

2.3. Visual Prompting

Visual prompting frames an input image as an *instruction* for a VLM, functioning similarly to how text prompts guide language models [30, 38, 59, 67, 69, 73, 80]. Numerous studies have demonstrated its effectiveness by exploiting the inherent image comprehension capabilities of VLMs. Set-of-Marks [71] augments each object in an image with its corresponding segmentation mask for more fine-grained visual grounding. Visual Sketchpad [22] provides tool-based framework for VLMs to utilize drawing tools to annotate images for complex tasks such as math problem solving and visual search. Recent research further proposes *visual* chain-of-thought (CoT) pipelines [9, 33, 54, 56, 68, 79, 80] that visualize intermediate reasoning steps as images and feed them back to the model as auxiliary inputs. This visual feedback loop has proven effective for spatial tasks, as it anchors textual reasoning to concrete visual cues [33, 68]. Building on this idea, we propose to transform an abstraction of a given scene and feed it back to a VLM in the form of a visual prompt, offering a new way for the model to reason from *arbitrary* viewpoints.

3. Method: Abstract Perspective Change

Our goal is to enable VLMs to solve spatial reasoning tasks from any given perspective (Fig. 2). Let us call the entity of the target perspective as the *reference viewer*. Since VLMs inherently approach spatial reasoning from an egocentric perspective [78], we propose to reformulate perspective-specific questions to align with the reference viewer's egocentric perspective. Inspired by theories in mental imagery [17, 48, 57], we begin by explicitly building an *abstraction* of the scene and use it as a foundation for shifting perspectives (Fig. 3).

Overview of `APC`. We call our approach Abstract Perspective Change (`APC`), which consists of three main stages.

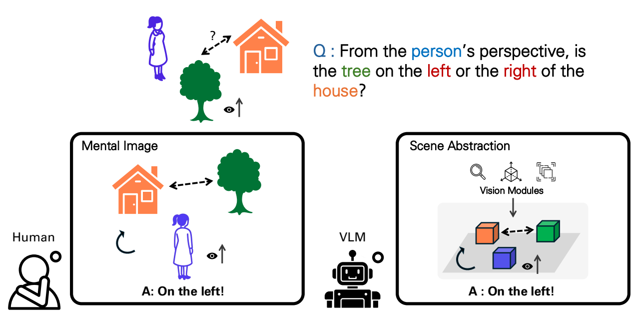

Figure 3. **Mental Imagery Simulation.** Inspired by how humans employ mental imagery to reason from across different perspectives (left), we propose a similar process for VLMs, by constructing an explicit abstraction of the input scene and using it as a foundation for perspective changes (right).

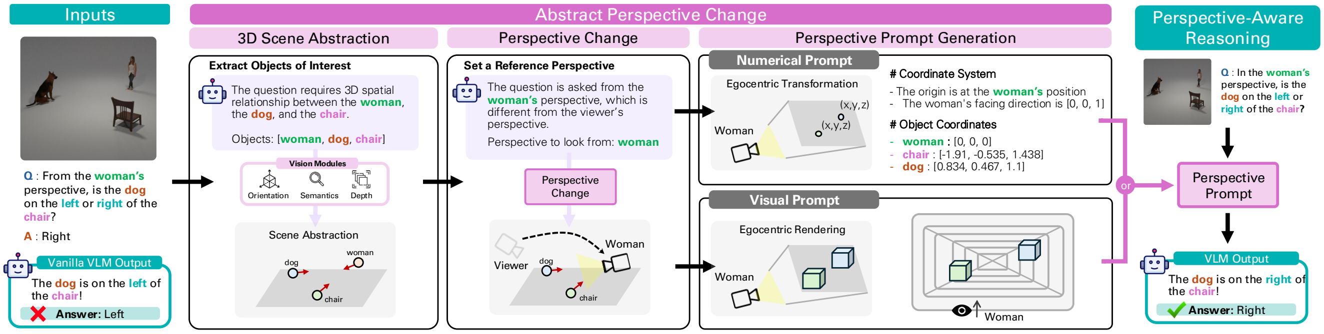

Figure 4. **Pipeline Overview of APC.** Our proposed framework consists of three stages. 1) Scene Abstraction (Sec. 3.1): APC first detects the objects of interest and build a coarse 3D abstraction of the scene using off-the-shelf vision foundation models. 2) Perspective Change (Sec. 3.2): Then, a reference perspective is set and the abstraction is transformed into the reference viewer's egocentric coordinate frame. 3) Perspective Prompting (Sec. 3.3): Finally, APC passes the transformed scene to the VLM by producing (1) a numerical (textual) prompt or (2) an abstract visual prompt, and poses the question of interest from the reference perspective.

(1) First, APC constructs a coarse 3D abstraction of the scene from the input image by selecting and extracting objects of interest using off-the-shelf vision modules (Sec. 3.1), drawing inspiration from human mental imagery [17]. (2) Next, APC selects a reference viewer for the spatial reasoning task among the objects of interests in the constructed scene abstraction. This determines "where to look from". Such a formulation allows the conversion of the allocentric reasoning problem to an egocentric spatial reasoning task by performing a *perspective change* that transforms the base coordinate system of the abstraction from the original camera view to that of the reference viewer (Sec. 3.2). (3) Finally, the transformed abstracted scene, which can now be posed as an egocentric problem, is fed back into the VLM for spatial reasoning (Sec. 3.3). We explore two alternative representations when providing the VLM with transformed astract scene information: 1) directly feeding numerical 3D coordinates of each object as a text prompt (numerical prompt), and 2) generating an abstract rendering of the scene as viewed by the reference perspective (visual prompt). An illustration of our APC pipeline is shown in Fig. 4, and we detail each step as follows.

3.1. Scene Abstraction

APC begins by building a coarse 3D abstraction of the scene. Given an image I and a spatial reasoning question Q, we define the abstraction of a scene as the set $S_E := \{O_i\}_{i=1}^n$ composed of objects of interest from the question Q. Here, E denotes that the abstraction is defined in the camera's egocentric coordinate system, and the number of objects of interest n is determined by the VLM based on Q. Each O_i corresponds to an object of interest in the image and is represented as a tuple (t_i, c_i, p_i), where t_i is the object's description, $c_i \in \mathbb{R}^3$ is its 3D position, and $p_i \in \mathbb{S}^3$ is a unit vector that indicates its orientation. Additionally, the camera is also included as an object of interest. This abstraction provides a minimal yet sufficient information in order to perform perspective changes, and mirrors how humans draw and rotate mental images when reasoning with perspectives [17, 57]. It allows for our APC to convert an allocentric problem to an egocentric spatial reasoning task, which VLMs can better solve [78]. More details are described below.

Extracting Objects of Interest. To determine which objects in the image should be included in the scene abstraction S_E, we provide the image I and the question Q to the VLM and instruct it to identify the list of objects necessary for answering the question. The VLM then returns the list of objects of interest, specified by their name, which we denote as t_i. The detailed instruction prompts are included in the **Appendix (Sec. E)**.

Building Object Abstractions. Given the list of objects of interest, we complete our abstracted scene representation by extracting the position and orientation of each object O_i using off-the-shelf vision foundation models. To obtain the 3D position of O_i, we first query GroundingDINO [43] with image I and the object description t_i and obtain its 2D bounding box b_i. We then crop I with b_i, and utilize SAM [28] to obtain a precise segmentation mask for O_i. Next, we extract the metric depth map of I using DepthPro [4] and unproject the pixels within the segmentation mask to 3D. Subsequently, the position c_i is obtained by taking the median coordinate of this 3D point cloud. For further implementation details, please refer to the **Appendix (Sec. C)**. Estimating the orientation p_i for each object O_i is also necessary to perform the desired perspective transformation. We utilize OrientAnything [66], which returns the object's frontal orientation within the camera co-

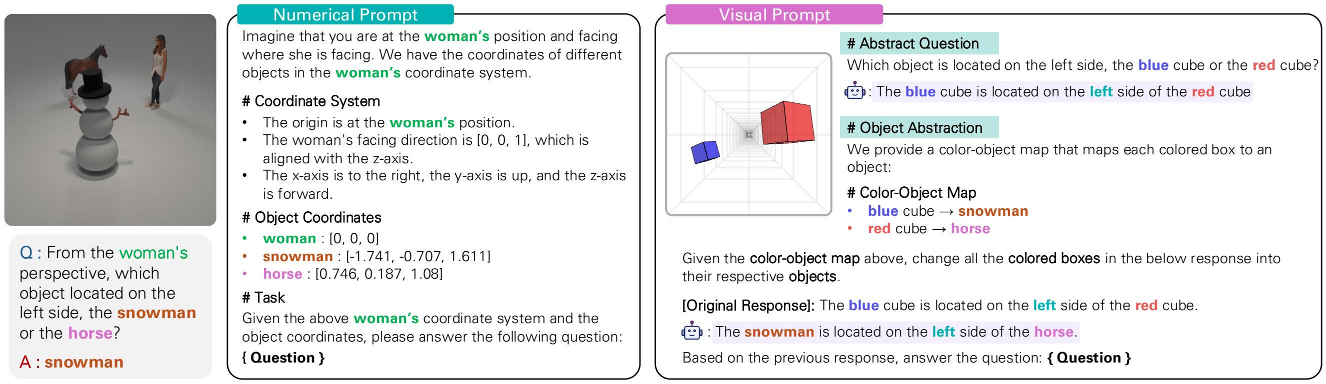

Figure 5. **Perspective Prompt Samples.** We explore two variations of perspective prompting, numerical (left) and visual (right). Numerical (textual) prompt is generated by directly utilizing the 3D coordinate and orientation information. To generate the Visual prompt, we first place a colored cube at each object's identified 3D position then render the scene at the reference viewpoint, which results in an egocentric depiction of the scene. In addition, we construct an abstract question along with object-color mapping to ground the abstracted view.

ordinate system. For this, we crop the image with b_i and feed the cropped image to OrientAnything to obtain O_i's orientation, hence completing our scene abstraction representation.

3.2. Perspective Change

With egocentric scene abstraction S_E for a given image I and question Q, APC then determines the reference viewer and performs perspective change to obtain a transformed scene abstraction from the reference viewer's perspective. This effectively converts an allocentric problem into an egocentric task, which VLMs find easier to handle.

Setting a Reference Perspective. APC first determines "where to look from" by selecting a reference viewer from the set of objects of interest. For this, we provide the spatial reasoning question Q to the VLM and instruct it to identify the reference perspective from which the question should be answered. We denote the extracted reference perspective as A, and provide the complete instruction for perspective extraction in the **Appendix (Sec. E)**.

Transforming Scene Abstraction. After identifying the reference viewer, we then transform the original camera-based scene abstraction S_E into the reference viewer's egocentric coordinate system. Specifically, we apply coordinate transformation from the camera's frame to that of the reference viewer A. In the resulting abstraction S_A, the reference viewer A is placed at the origin, and its orientation is aligned with the z-axis. This step supports APC's main objective of reframing a general perspective question—typically an allocentric problem—into the reference viewer's egocentric viewpoint, making it an egocentric task. Finally, we provide S_A to the VLM so it can answer the question Q from A's perspective. We describe this stage more in depth below.

3.3. Perspective Prompting

The final step of APC involves generating a prompt from the transformed scene abstraction S_A to feed as input for the VLM. That is, how is the VLM asked with the transformed, now egocentric spatial reasoning task? We refer to our generated prompt as the *perspective prompt* for image I and question Q. Since VLMs can take images and text inputs, we explore two choices for the representation of this prompt: numerical (textual) and visual.

Numerical (Textual) Prompt. Recall that an object abstraction in the transformed scene abstraction S_A consists of the object's textual description, its corresponding 3D position, and its orientation, *i.e.* $O'_i = (t_i, c'_i, p'_i)$. Hence, a straightforward approach is to directly feed this information into the VLM. Specifically, we include the 3D position c'_i in a predefined instruction template and instruct the VLM to directly solve the question Q. The full instruction template is provided in the **Appendix (Sec. E)**.

Visual Prompt. Our goal is to let VLMs "view the scene from A's perspective"; thus an alternative choice for the perspective prompt is a visualization of our abstraction S_A. We begin by assigning each object an equal-sized cube, with each cube's position matching the objects' positions c'_i. We then render these cubes from the reference viewer A's vantage point, generating an egocentric depiction of the scene abstraction. To distinguish between objects, each cube is assigned a unique color. When providing this information to the VLM, we modify the original question Q to reflect the abstract visual representation. Specifically, we replace object names (*e.g.* *"dog"*) with their corresponding colored cubes (*e.g.* *"red box"*), forming an abstract question Q^*. Refer to Fig. 5 for an example of an obtained abstraction question. Putting it all together, the VLM receives as a prompt both the abstract rendered image—showing colored

ICCV #17

cubes—and the reformulated question Q^*. This allows it to answer spatial reasoning questions originally posed from *arbitrary* perspectives by reasoning with this abstract, egocentric visual prompt. More details on the visual prompting process are presented in the **Appendix (Sec. C.3)**.

4. Results

In this section, we present the experimental results of our APC across a range of spatial reasoning tasks that include specified reference perspectives. We compare APC to multiple baseline methods and show how our abstraction-based allocentric-to-egocentric reasoning framework enables the VLM to handle alternative perspectives. We use Qwen2.5-VL [2] as our backbone VLM.

4.1. Evaluation Settings

Benchmarks. We validate our APC on both synthetic [78] and real-world [46] benchmarks in which the spatial reasoning requires perspective changes. Sample image-question pairs from each benchmark are shown in Fig. 6.

- **COMFORT++:** Zhang *et al*. [78] introduce COMFORT, a benchmark synthesis protocol designed to evaluate VLMs on perspective-aware spatial reasoning. It employs a simple Blender [13] rendering pipeline to place multiple objects in a synthetic scene with one reference viewer and various other objects. Each scene poses a spatial reasoning question from the reference viewer's perspective. Building on COMFORT, we construct four types of spatial reasoning tasks that require a reference viewer different from the camera: *left/right*, *closer/further*, *visibility*, and *facing*.
- **3DSRBench:** Ma *et al*. [46] introduce a 3D spatial reasoning benchmark based on MS-COCO images [36]. We focus on three categories that require an allocentric viewpoint: *left/right*, *visibility*, and *facing*. Note that we recast the original *front/behind* question in 3DSRBench into a *visibility* question using the same images. We provide further discussion on the dataset and the evaluation protocol in the **Appendix (Sec. D)**.

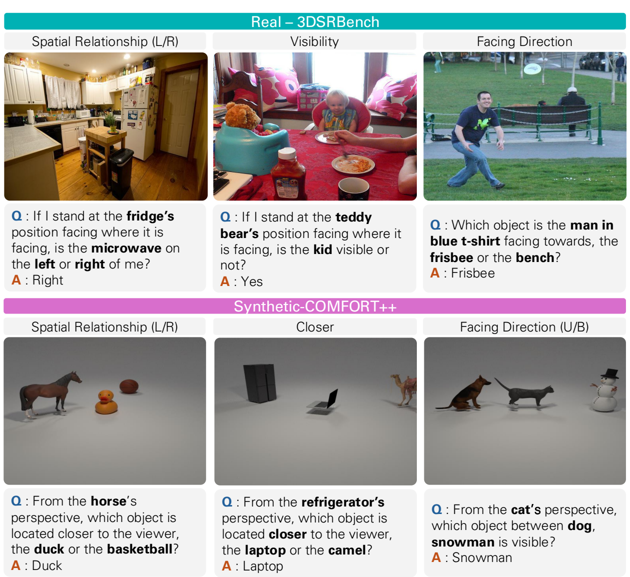

Figure 6. **Benchmark Visualization.** Example image-question pairs from 3DSRBench [46] and COMFORT++ [78] benchmarks. The tasks probe spatial reasoning across left-right relations, object visibility, closenss, and the facing direction of objects.

Baselines: VLMs. We benchmark our APC against multiple state-of-the-art VLMs, including both open-source and proprietary models. For open-source, we include LLaVA-NeXT [41], LLaVA-OneVision [31], Molmo [14], and Qwen2.5-VL [2]. We also include proprietary models: GPT-4o [24] and Gemini-2.0-Flash [63]. We refer to these as *pure VLMs*. Additionally, we compare against *grounded VLMs*, which include models explicitly tuned for spatial reasoning, such as SpatialVLM [8] and SpatialRGPT [10]. We also include SpatialPIN [45], which leverages interactions between VLMs and vision foundation models for complex spatial reasoning.

Baselines: Dense Reconstruction. To compare APC with standard dense reconstruction techniques for novel view synthesis, we introduce two baselines. First, we extend SpatialPIN [45] to include our perspective change phase (Sec. 3.2). We use the generated meshes from its original pipeline and render the meshes from the reference perspective, and denote this extension as SpatialPIN*. Refer to the **Appendix (Sec. B)** for more details. Second, we adopt ViewCrafter [75], a novel view synthesis method designed for single-image inputs. For both baselines, we synthesize a novel view according to the reference perspective's relative pose, and feed the resulting image to the VLM for spatial reasoning.

4.2. Evaluation on COMFORT++ [78]

Tab. 1 (cols 2-5) provides quantitative comparisons on COMFORT++. Here, APC-Vis refers to our visual prompt, and APC-Num corresponds to the numerical prompt. Even though the benchmark consists of objects rendered in a simple, synthetic scene (see Fig. 6), we find that most pure VLMs (rows 3-9) struggle with the *left/right* task, hovering around chance level with the best performing model LLaVA-OneVision scoring only 55.33%. This confirms earlier observations [78] that VLMs fail to adopt alternative perspectives. Even specialist VLMs designed for spatial reasoning perform poorly, with SpatialVLM at 46.0% and both SpatialRGPT and SpatialPIN also exhibiting low accuracy. We observed that SpatialRGPT often generates hallucinated responses unrelated to the instruction,

Method	COMFORT++ [78]				3DSRBench [46]		
	left/right	closer	visibility	facing	left/right	visibility	facing
Random	50.00	50.00	50.00	50.00	50.00	50.00	50.00
LLaVA-NeXT [41]	48.00	47.33	40.00	39.00	34.10	41.57	50.29
LLaVA-OneVision [31]	55.33	79.00	50.94	38.33	32.09	46.51	60.12
Molmo [14]	36.33	35.67	31.88	29.00	19.77	22.97	32.08
Qwen2.5-VL [2]	43.33	74.33	51.25	43.00	34.96	45.06	53.47
Cambrian-1 [64]	52.00	79.00	57.50	41.00	40.97	49.71	65.03
GPT-4o [24]	41.00	61.33	53.75	38.67	2.01	40.12	47.70
Gemini-2.0-Flash [63]	43.67	26.00	40.31	13.00	24.93	57.65	55.20
SpatialVLM [8]	46.00	41.67	42.81	29.33	22.35	46.51	47.11
SpatialRGPT [10]	27.08	33.90	29.25	1.33	25.98	27.19	42.55
SpatialPIN [45]	19.62	23.96	48.43	43.91	11.10	42.40	11.66
SpatialPIN* [45]	59.80	70.45	49.84	50.51	50.10	52.30	28.86
ViewCrafter [75]	32.33	53.00	38.75	37.46	28.41	22.47	18.31
APC-Num (Ours)	88.67	**96.00**	71.25	62.00	71.92	62.79	60.98
APC-Vis (Ours)	**89.67**	94.33	**90.00**	**88.33**	**72.78**	**67.44**	**66.47**

Table 1. **Quantitative Comparisons.** Purple ( ) represents *pure VLMs*, green ( ) represents *grounded VLMs*, and red ( ) represents *dense reconstruction-based* frameworks. Gray ( ) corresponds to our APC. **Bold** and underline indicate the best and the second-best result for each column, respectively. APC-Num and APC-Vis refer to our method employing numerical prompt and visual prompt, respectively.

thereby resulting in low accuracy. While SpatialPIN*—employing perspective change—shows better performance, low-quality meshes often bottleneck further improvements (refer to Sec. B for more discussions). In contrast, our APC significantly outperforms these baselines, achieving 89.67% accuracy with a visual prompt and 88.67% with a numerical (textual) prompt.

For the *closer* task, some VLMs show relatively high accuracy (79.00% for both LLaVA-OneVision and Cambrian-1), likely since they can also solve the question by comparing object distances directly from the egocentric viewpoint. Even in this case, APC achieves higher accuracy, attaining 96% when using a numerical prompt. Moreover, for *visibility* and *facing* categories, the baseline models perform at near-chance levels, failing to take the reference perspectives into account. Notably, APC exhibits a performance gap between visual and numerical prompts, with the visual prompt outperforming the numerical one by +18.75% and +26.33%, respectively. We attribute this difference to trivial logical errors that VLMs often make when relying on numerical coordinates. In contrast, for these two tasks the visual prompt requires only simple visual perception, mitigating such logical errors and achieving more accurate results.

4.3. Evaluation on 3DSRBench [46]

Tab. 1 (cols 6-8) presents quantitative comparisons on 3DSRBench, which includes real images. Compared to the synthetic environment in COMFORT++, using real images introduces additional noise into both the VLMs' visual reasoning and our APC's scene abstraction phase. For the *left/right* task, baseline VLMs consistently fall under 50% accuracy, including the grounded VLMs. Even SpatialPIN* with perspective change only reaches 50.10%, at near chance-level. We find that using ViewCrafter to generate a novel view from the reference perspective yields 28.41% accuracy, due to the noise and hallucinations during generation. We provide visualizations of the rendered views from both SpatialPIN* [45] and ViewCrafter [75] in the **Appendix (Sec. B)**, along with more discussions on the different between our abstraction-based approach and the dense reconstruction-based approaches. Compared to other baselines, APC consistently achieves accuracies above 60% both with visual and numerical prompts, showing that our framework is robust to real images.

For *visibility* task, while our method outperforms the baselines, the accuracy appears lower than previous tasks—67.44% with the visual prompt and 62.79% with the numerical prompt. We attribute this decline to the noise in the scene abstraction phase, particularly errors in detected orientations or centroids. This issue is evident when compared to the *visibility* task in COMFORT++, which employs the same question format but with a simpler scene. Lastly, for the *facing* task, Cambrian-1 achieves 64.03%, yet our APC with a visual prompt still leads at 66.47%. Qualitative examples on 3DSRBench are shown in Fig. 7.

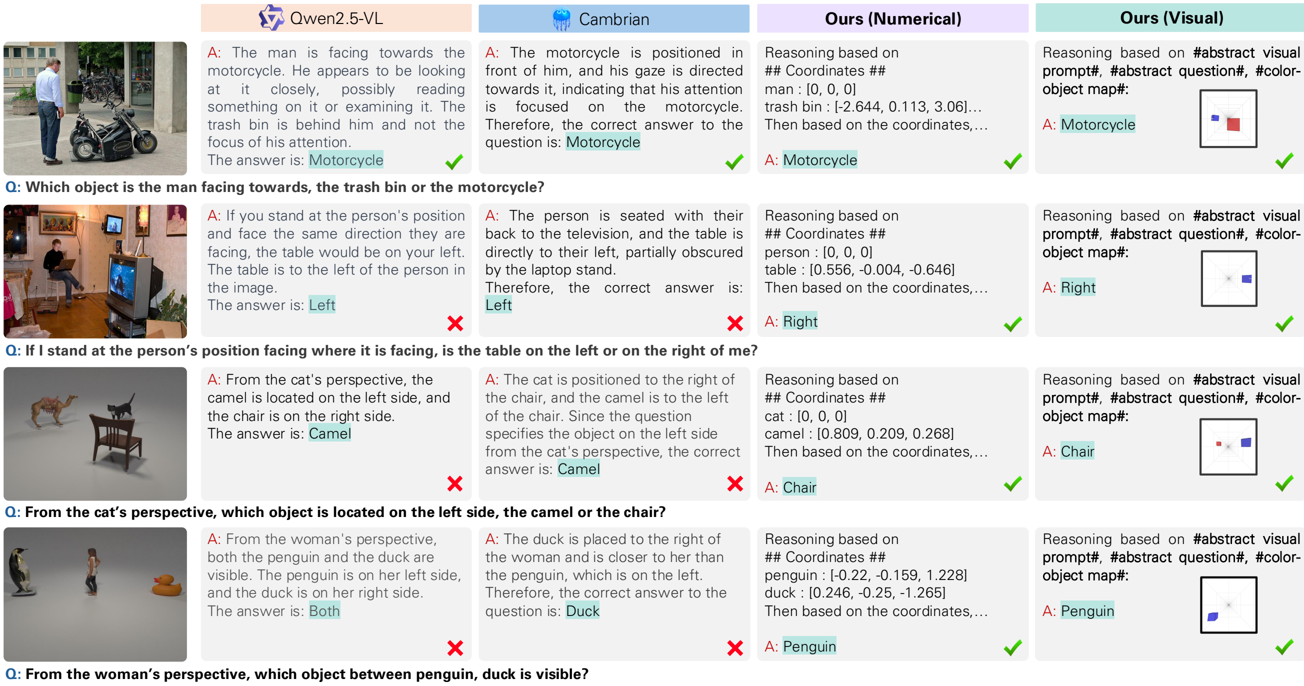

Figure 7. **Spatial Reasoning with Perspective Change.** Recent VLMs such as Qwen2.5-VL [2] and Cambrian-1 [64] often struggle with spatial reasoning tasks that require a shift to a specific reference viewpoint. In contrast, our APC effectively handles such perspective changes by constructing a scene abstraction and delivering the transformed view through a simple prompting technique.

4.4. Probing the Perspective Awareness of VLMs

Finally, we analyze the *perspective-awareness* of each method by assessing spatial reasoning accuracy across different viewpoints. Specifically, we select two tasks—*left/right* and *closer*—and construct 60 scenes similar following our setting in COMFORT++. Each scene is rendered from 20 evenly spaced azimuth angles. We define θ as the angular offset between the camera's orientation and the reference viewer's orientation. Here, for $\theta = 0°$ the camera is aligned with the reference perspective, while $\theta = 180°$ indicates that the reference viewer is facing towards the camera.

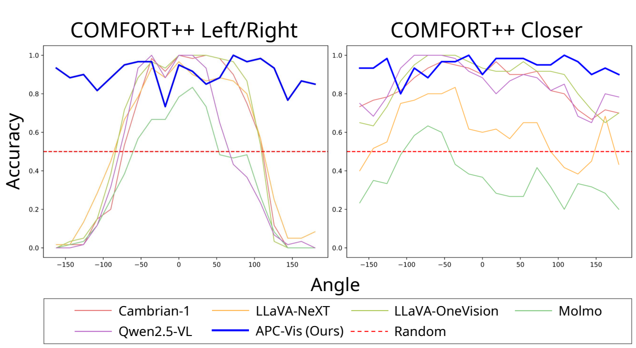

Figure 8. **Perspective Awareness.** Each plot shows accuracy versus the angular offset θ between the camera and the reference viewpoint. While baselines show clear degradation at certain ranges of θ, APC retains robust accuracy across all angles, demonstrating strong perspective-aware reasoning.

The results are shown in Fig. 8. For the *left/right* task (left), the baselines exhibit clear bell-shaped curves, achieving near perfect accuracy when θ is close to $0°$ (egocentric) but rapidly declining as the magnitude of θ increases (allocentric). In contrast, APC maintains consistently high accuracy across all angles, demonstrating strong *perspective-aware* reasoning. For the *closer* task (right), baseline models also show noticeable accuracy drops, especially near the leftmost and rightmost θ ranges. APC consistently achieves over 80% accuracy, robustly handling viewpoints regardless of their deviation from the egocentric perspective.

5. Conclusion

In this work, we introduced APC, a framework empowering VLMs with the capability of perspective-aware reasoning. Our key idea is to simulate the mental imagery process of humans, abstracting the scene in an image to facilitate allocentric-to-egocentric perspective shifts, and in turn convey the transformed view to the VLM in the form of a prompt. The scene abstraction is constructed using vision foundation models for object detection, segmentation, and orientation estimation. The reframed prompt from the new perspective, either in text or image form, is then processed by VLMs, leveraging their egocentric reasoning capabilities. As shown by our experiment on both synthetic and real spatial reasoning benchmarks, APC enables robust accuracy across diverse perspectives, thereby opening new possibilities of VLMs on real-world spatial tasks.

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. In *NeurIPS*, 2022. 1

[2] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 1, 3, 6, 7, 8

[3] 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*, 2023. 3

[4] 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 *ICLR*, 2025. 4

[5] Neil Burgess. Spatial memory: how egocentric and allocentric combine. *Trends in cognitive sciences*, 2006. 2

[6] Wenxiao Cai, Iaroslav Ponomarenko, Jianhao Yuan, Xiaoqi Li, Wankou Yang, Hao Dong, and Bo Zhao. Spatialbot: Precise spatial understanding with vision language models. In *ICRA*, 2025. 2, 3

[7] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, 2020. 2

[8] Boyuan Chen, Zhuo Xu, Sean Kirmani, Brain Ichter, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. In *CVPR*, 2024. 1, 2, 3, 6, 7

[9] Zhenfang Chen, Qinhong Zhou, Yikang Shen, Yining Hong, Zhiqing Sun, Dan Gutfreund, and Chuang Gan. Visual chain-of-thought prompting for knowledge-based visual reasoning. In *AAAI*, 2024. 3

[10] 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 model. In *NeurIPS*, 2024. 1, 2, 3, 6, 7

[11] Herbert H. Clark and Susan E. Brennan. Grounding in communication. *Perspectives on Socially Shared Cognition*, 1991. 3

[12] Geoff G Cole, Steven Samuel, and Madeline J Eacott. A return of mental imagery: The pictorial theory of visual perspective-taking. *Consciousness and Cognition*, 2022. 2

[13] Blender Online Community. Blender - a 3d modelling and rendering package, 2018. 6

[14] 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 multimodal models. *arXiv preprint arXiv:2409.17146*, 2024. 1, 3, 6, 7

[15] Vaibhav A Diwadkar and Timothy P McNamara. Viewpoint dependence in scene recognition. *Psychological science*, 1997. 2

[16] Linxi Fan, Guanzhi Wang, Yunfan Jiang, Ajay Mandlekar, Yuncong Yang, Haoyi Zhu, Andrew Tang, De-An Huang, Yuke Zhu, and Anima Anandkumar. Minedojo: Building open-ended embodied agents with internet-scale knowledge. In *NeurIPS*, 2022. 2

[17] RA Finke. Principles of mental imagery, 1989. 2, 3, 4

[18] Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A Smith, Wei-Chiu Ma, and Ranjay Krishna. Blink: Multimodal large language models can see but not perceive. In *ECCV*, 2024. 3

[19] Gracjan Góral, Alicja Ziarko, Michal Nauman, and Maciej Wołczyk. Seeing through their eyes: Evaluating visual perspective taking in vision language models. *arXiv preprint arXiv:2409.12969*, 2024. 2, 3

[20] Tanmay Gupta and Aniruddha Kembhavi. Visual programming: Compositional visual reasoning without training. In *CVPR*, 2023. 3

[21] Miran Heo, Min-Hung Chen, De-An Huang, Sifei Liu, Subhashree Radhakrishnan, Seon Joo Kim, Yu-Chiang Frank Wang, and Ryo Hachiuma. Omni-rgpt: Unifying image and video region-level understanding via token marks. In *CVPR*, 2025. 2, 3

[22] Yushi Hu, Weijia Shi, Xingyu Fu, Dan Roth, Mari Ostendorf, Luke Zettlemoyer, Noah A Smith, and Ranjay Krishna. Visual sketchpad: Sketching as a visual chain of thought for multimodal language models. In *NeurIPS*, 2024. 3

[23] Yushi Hu, Otilia Stretcu, Chun-Ta Lu, Krishnamurthy Viswanathan, Kenji Hata, Enming Luo, Ranjay Krishna, and Ariel Fuxman. Visual program distillation: Distilling tools and programmatic reasoning into vision-language models. In *CVPR*, 2025. 1

[24] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024. 1, 3, 6, 7

[25] Linyi Jin, Jianming Zhang, Yannick Hold-Geoffroy, Oliver Wang, Kevin Blackburn-Matzen, Matthew Sticha, and David F Fouhey. Perspective fields for single image camera calibration. In *CVPR*, 2023. 3

[26] Amita Kamath, Jack Hessel, and Kai-Wei Chang. What's "up" with vision-language models? investigating their struggle with spatial reasoning. In *EMNLP*, 2023. 3

[27] 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. In *ICCV*, 2023. 2

[28] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *ICCV*, 2023. 4

[29] S. M. Kosslyn, T. M. Ball, and B. J. Reiser. Visual images preserve metric spatial information: Evidence from studies of image scanning. *Journal of Experimental Psychology: Human Perception and Performance*, 1978. 2

[30] Xuanyu Lei, Zonghan Yang, Xinrui Chen, Peng Li, and Yang Liu. Scaffolding coordinates to promote vision-language coordination in large multi-modal models. *arXiv preprint arXiv:2402.12058*, 2024. 3

[31] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024. 2, 3, 6, 7

[32] Chengzu Li, Caiqi Zhang, Han Zhou, Nigel Collier, Anna Korhonen, and Ivan Vulić. Topviewrs: Vision-language models as top-view spatial reasoners. In *EMNLP*, 2024. 3

[33] Chengzu Li, Wenshan Wu, Huanyu Zhang, Yan Xia, Shaoguang Mao, Li Dong, Ivan Vulić, and Furu Wei. Imagine while reasoning in space: Multimodal visualization-of-thought. *arXiv preprint arXiv:2501.07542*, 2025. 3

[34] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, 2022. 1

[35] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*, 2023. 1

[36] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 6

[37] Drew Linsley, Peisen Zhou, Alekh Karkada Ashok, Akash Nagaraj, Gaurav Gaonkar, Francis E Lewis, Zygmunt Pizlo, and Thomas Serre. The 3d-pc: a benchmark for visual perspective taking in humans and machines. In *ICLR*, 2025. 2, 3

[38] Benlin Liu, Yuhao Dong, Yiqin Wang, Yongming Rao, Yansong Tang, Wei-Chiu Ma, and Ranjay Krishna. Coarse correspondence elicit 3d spacetime understanding in multimodal language model. In *CVPR*, 2025. 3

[39] Fangyu Liu, Guy Emerson, and Nigel Collier. Visual spatial reasoning. In *EMNLP*, 2023. 3

[40] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023. 1, 3

[41] 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. 3, 6, 7

[42] Minghua Liu, Ruoxi Shi, Linghao Chen, Zhuoyang Zhang, Chao Xu, Xinyue Wei, Hansheng Chen, Chong Zeng, Jiayuan Gu, and Hao Su. One-2-3-45++: Fast single image to 3d objects with consistent multi-view generation and 3d diffusion. In *CVPR*, 2024. 3

[43] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *ECCV*, 2024. 2, 4

[44] Yuecheng Liu, Dafeng Chi, Shiguang Wu, Zhanguang Zhang, Yaochen Hu, Lingfeng Zhang, Yingxue Zhang, Shuang Wu, Tongtong Cao, Guowei Huang, et al. Spatialcot: Advancing spatial reasoning through coordinate alignment and chain-of-thought for embodied task planning. *arXiv preprint arXiv:2501.10074*, 2025. 3

[45] Chenyang Ma, Kai Lu, Ta-Ying Cheng, Niki Trigoni, and Andrew Markham. Spatialpin: Enhancing spatial reasoning capabilities of vision-language models through prompting and interacting 3d priors. In *NeurIPS*, 2024. 1, 2, 3, 6, 7

[46] Wufei Ma, Haoyu Chen, Guofeng Zhang, Celso M de Melo, Alan Yuille, and Jieneng Chen. 3dsrbench: A comprehensive 3d spatial reasoning benchmark. *arXiv preprint arXiv:2412.07825*, 2024. 2, 3, 6, 7

[47] Damiano Marsili, Rohun Agrawal, Yisong Yue, and Georgia Gkioxari. Visual agentic ai for spatial reasoning with a dynamic api. In *CVPR*, 2025. 3

[48] Bence Nanay. Mental imagery. *The Stanford Encyclopedia of Philosophy*, 2021. 2, 3

[49] A. Paivio. *Imagery and Verbal Processes (1st ed.)*. Psychology Press, 1979. 2

[50] Pooyan Rahmanzadehgervi, Logan Bolton, Mohammad Reza Taesiri, and Anh Totti Nguyen. Vision language models are blind. In *ACCV*, 2024. 3

[51] Santhosh Kumar Ramakrishnan, Erik Wijmans, Philipp Kraehenbuehl, and Vladlen Koltun. Does spatial cognition emerge in frontier models? In *ICLR*, 2025. 3

[52] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. Sam 2: Segment anything in images and videos. *arXiv*, 2024. 2

[53] Arijit Ray, Jiafei Duan, Reuben Tan, Dina Bashkirova, Rose Hendrix, Kiana Ehsani, Aniruddha Kembhavi, Bryan A Plummer, Ranjay Krishna, Kuo-Hao Zeng, et al. Sat: Spatial aptitude training for multimodal language models. *arXiv preprint arXiv:2412.07755*, 2024. 3

[54] Daniel Rose, Vaishnavi Himakunthala, Andy Ouyang, Ryan He, Alex Mei, Yujie Lu, Michael Saxon, Chinmay Sonar, Diba Mirza, and William Yang Wang. Visual chain of thought: bridging logical gaps with multimodal infillings. *arXiv preprint arXiv:2305.02317*, 2023. 3

[55] Paola Del Sette, Markus Bindemann, and Heather J Ferguson. Visual perspective-taking in complex natural scenes. *Quarterly Journal of Experimental Psychology*, 2022. 3

[56] Hao Shao, Shengju Qian, Han Xiao, Guanglu Song, Zhuofan Zong, Letian Wang, Yu Liu, and Hongsheng Li. Visual cot: Advancing multi-modal language models with a comprehensive dataset and benchmark for chain-of-thought reasoning. *NeurIPS*, 2024. 3

[57] Roger N Shepard and Jacqueline Metzler. Mental rotation of three-dimensional objects. *Science*, 171(3972):701–703, 1971. 2, 3, 4

[58] Fatemeh Shiri, Xiao-Yu Guo, Mona Far, Xin Yu, Reza Haf, and Yuan-Fang Li. An empirical analysis on spatial reasoning capabilities of large multimodal models. In *EMNLP*, 2024. 3

[59] Aleksandar Shtedritski, Christian Rupprecht, and Andrea Vedaldi. What does clip know about a red circle? visual prompt engineering for vlms. In *ICCV*, 2023. 3

[60] Chan Hee Song, Valts Blukis, Jonathan Tremblay, Stephen Tyree, Yu Su, and Stan Birchfield. Robospatial: Teaching spatial understanding to 2d and 3d vision-language models for robotics. In *CVPR*, 2025. 3

[61] Dídac Surís, Sachit Menon, and Carl Vondrick. Vipergpt: Visual inference via python execution for reasoning. In *ICCV*, 2023. 3

[62] Yihong Tang, Ao Qu, Zhaokai Wang, Dingyi Zhuang, Zhaofeng Wu, Wei Ma, Shenhao Wang, Yunhan Zheng, Zhan Zhao, and Jinhua Zhao. Sparkle: Mastering basic spatial capabilities in vision language models elicits generalization to composite spatial reasoning. *arXiv preprint arXiv:2410.16162*, 2024. 3

[63] Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024. 1, 3, 6, 7

[64] 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. In *NeurIPS*, 2024. 1, 2, 3, 7, 8

[65] Jiayu Wang, Yifei Ming, Zhenmei Shi, Vibhav Vineet, Xin Wang, Sharon Li, and Neel Joshi. Is a picture worth a thousand words? delving into spatial reasoning for vision language models. In *NeurIPS*, 2024. 3

[66] Zehan Wang, Ziang Zhang, Tianyu Pang, Chao Du, Hengshuang Zhao, and Zhou Zhao. Orient anything: Learning robust object orientation estimation from rendering 3d models. *arXiv preprint arXiv:2412.18605*, 2024. 2, 4

[67] Junda Wu, Zhehao Zhang, Yu Xia, Xintong Li, Zhaoyang Xia, Aaron Chang, Tong Yu, Sungchul Kim, Ryan A Rossi, Ruiyi Zhang, et al. Visual prompting in multimodal large language models: A survey. *arXiv preprint arXiv:2409.15310*, 2024. 3

[68] Wenshan Wu, Shaoguang Mao, Yadong Zhang, Yan Xia, Li Dong, Lei Cui, and Furu Wei. Mind's eye of llms: Visualization-of-thought elicits spatial reasoning in large language models. In *NeurIPS*, 2024. 3

[69] Yixuan Wu, Yizhou Wang, Shixiang Tang, Wenhao Wu, Tong He, Wanli Ouyang, Philip Torr, and Jian Wu. Dettoolchain: A new prompting paradigm to unleash detection ability of mllm. In *ECCV*, 2024. 3

[70] 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. In *NeurIPS*, 2023. 2

[71] Jianwei Yang, Hao Zhang, Feng Li, Xueyan Zou, Chunyuan Li, and Jianfeng Gao. Set-of-mark prompting unleashes extraordinary visual grounding in gpt-4v. *arXiv preprint arXiv:2310.11441*, 2023. 3

[72] Jihan Yang, Shusheng Yang, Anjali W Gupta, Rilyn Han, Li Fei-Fei, and Saining Xie. Thinking in space: How multimodal large language models see, remember, and recall spaces. In *CVPR*, 2025. 3

[73] Lingfeng Yang, Yueze Wang, Xiang Li, Xinlong Wang, and Jian Yang. Fine-grained visual prompting. In *NeurIPS*, 2023. 3

[74] 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 *ICRA*, 2025. 2

[75] Wangbo Yu, Jinbo Xing, Li Yuan, Wenbo Hu, Xiaoyu Li, Zhipeng Huang, Xiangjun Gao, Tien-Tsin Wong, Ying Shan, and Yonghong Tian. Viewcrafter: Taming video diffusion models for high-fidelity novel view synthesis. *arXiv preprint arXiv:2409.02048*, 2024. 6, 7

[76] Wentao Yuan, Jiafei Duan, Valts Blukis, Wilbert Pumacay, Ranjay Krishna, Adithyavairavan Murali, Arsalan Mousavian, and Dieter Fox. Robopoint: A vision-language model for spatial affordance prediction for robotics. In *CORL*, 2024. 3

[77] Wenyu Zhang, Wei En Ng, Lixin Ma, Yuwen Wang, Jungqi Zhao, Boyang Li, and Lu Wang. Sphere: A hierarchical evaluation on spatial perception and reasoning for vision-language models. *arXiv preprint arXiv:2412.12693*, 2024. 2, 3

[78] 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 *ICLR*, 2025. 2, 3, 4, 6, 7

[79] Qingqing Zhao, Yao Lu, Moo Jin Kim, Zipeng Fu, Zhuoyang Zhang, Yecheng Wu, Zhaoshuo Li, Qianli Ma, Song Han, Chelsea Finn, et al. Cot-vla: Visual chain-of-thought reasoning for vision-language-action models. In *CVPR*, 2025. 3

[80] Qiji Zhou, Ruochen Zhou, Zike Hu, Panzhong Lu, Siyang Gao, and Yue Zhang. Image-of-thought prompting for visual reasoning refinement in multimodal large language models. *arXiv preprint arXiv:2405.13872*, 2024. 3

[81] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. In *ICLR*, 2021. 2