Spatial Intelligence in Vision-Language Models: A Comprehensive Survey

Disheng Liu[1] **Tuo Liang**[1] **Zhe Hu**[1] **Jierui Peng**[1] **Yiren Lu**[1]
Yi Xu[2] **Yun Fu**[2] **Yu Yin**[1†]

[1]Department of Computer and Data Sciences, Case Western Reserve University
[2]Department of Electrical and Computer Engineering, Northeastern University

Vision-Language Models (VLMs) have achieved remarkable success but exhibit a fundamental deficiency in spatial intelligence, a critical capability for progress in embodied AI, autonomous driving, and spatially coherent generation. In response, the research community has produced an explosion of work dedicated to enhancing these models, but this rapid progress has resulted in a fragmented and disorganized landscape lacking a unified framework. This paper presents the first comprehensive survey to address this gap, uniquely providing a systematic review that spans the foundations of spatial intelligence in VLMs, root causes of spatial limitations, enhancement methodologies, evaluation protocols, and real-world applications. Specifically, we introduce a novel, intervention-based taxonomy that categorizes enhancement methodologies according to where spatial information is incorporated: (1) training-free prompting, (2) model-centric enhancements (training strategies, architectural modules, encoder improvements), (3) explicit 2D information injection, (4) 3D spatial enrichment, and (5) data-centric approaches. To further assess the true capabilities of current models, we conduct a rigorous empirical study evaluating 37 models across 9 representative benchmarks. Our results and analysis reveal the state-of-the-art, identify the strengths and weaknesses of different methods, and uncover critical limitations in existing evaluation protocols. By structuring this rapidly evolving field and establishing a clear research agenda, this survey serves as an indispensable resource for advancing the next generation of spatially intelligent AI systems.

†: Corresponding Author

Keywords: Spatial Intelligence, Vision Language Models, Foundation Models, Multimodal Large Language Models, Spatial Reasoning

Date: October 30, 2025
Contact: disheng.liu@case.edu, yu.yin@case.edu
Github Repository: https://github.com/vulab-AI/Awesome-Spatial-VLMs
Evaluation Dataset: https://huggingface.co/datasets/LLDDSS/Awesome_Spatial_VQA_Benchmarks

Contents

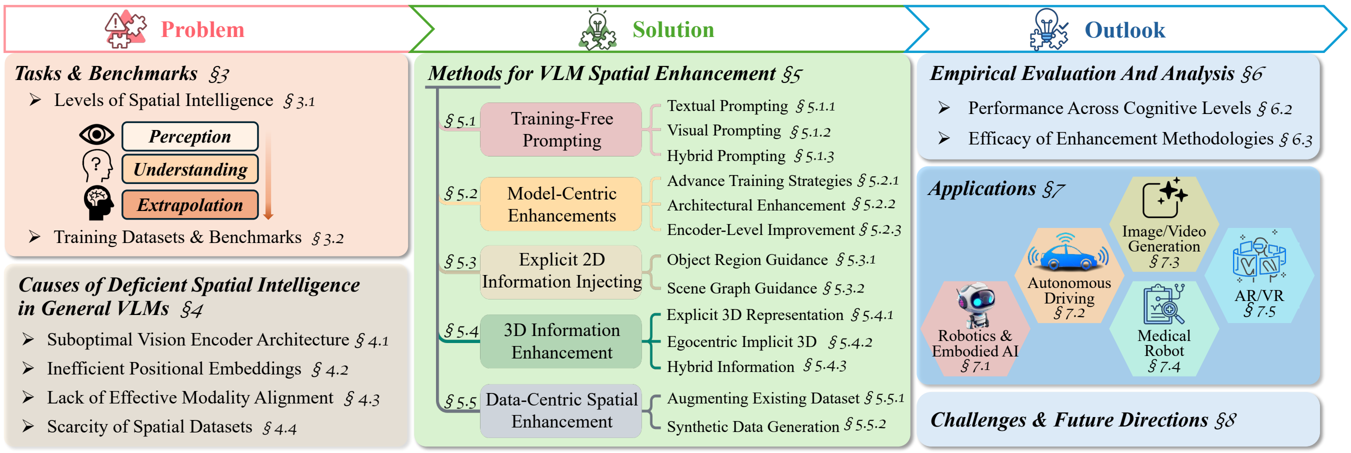

Figure 1 Survey outline. We first establish the foundations of spatial intelligence by defining a hierarchy of tasks and analyzing the root causes of model deficiencies (§ 3-§ 4). We then systematically review and categorize the diverse methods developed for spatial enhancement (§ 5), culminating in a large-scale empirical evaluation of representative models (§ 6), a review of key applications § 7), and a discussion of open challenges and future directions (§ 8).

1 Introduction

Vision-Language Models (VLMs) have achieved remarkable success in integrating visual and textual information, yet they consistently falter on a fundamental aspect of intelligence: **spatial reasoning**. While humans seamlessly combine perception, language, and memory to navigate and interact with their environment [1–3], the absence of robust spatial intelligence represents a critical bottleneck for the next generation of multimodal AI. This spatial deficiency prevents the deployment of multimodal AI in high-stakes domains. In embodied AI, 3D spatial awareness is the prerequisite for meaningful physical interaction [4]. In autonomous driving, understanding dynamic spatial relations is paramount for ensuring safety and reliability [5, 6]. Similarly, applications in augmented reality and generative AI depend on spatial coherence to produce immersive and physically plausible scenes [7–11].

The urgent need for spatial intelligence has triggered a rapid surge of research about developing spatially-aware VLMs. From 2023 to 2025, the field has progressed from early prototypes into a diverse, complex ecosystem at unprecedented speed. This acceleration is reflected in: 1) the rapid increase of distinct model families [12–16]; 2) the construction of large-scale training corpora with 46 millions of spatially grounded annotations [17–19]; 3) the introduction of 49 benchmarks to probe these new capabilities (Sec. 3; Appx. Tab. 3). While this rapid progress signifies a clear inflection point, it has also yielded a fragmented landscape of disparate methods, terminologies, and evaluation protocols, creating a critical need for consolidation.

Despite the prosperity in spatially-aware VLM research, the field lacks a unifying survey to structure its rapid growth. Existing reviews fall into two categories, neither providing a complete picture. First, broad surveys of foundation models largely omit spatial intelligence as a topic of focused analysis, prioritizing instead training paradigms and general applications [20, 21]. Second, the few spatially-related reviews that do exist are narrow in scope: Some concentrate exclusively on scene-level representations, like multi-view images and 3D representations like point clouds and meshes, treating spatial intelligence as a byproduct of 3D modeling [22, 23]; Others cover applications in an interdisciplinary manner without a systematic technical review [24]. Consequently, a comprehensive and technically-focused survey is urgently needed to unify the field, provide a rigorous evaluation of both general and specialized models, and establish a clear roadmap for future research.

To fill this blank, this paper presents the first comprehensive survey to systematically structure the field, encompassing the foundations of spatial intelligence in VLMs, the underlying causes of spatial limitations, enhancement methodologies, evaluation protocols, and real-world applications. We organize the field through the lens of **methodological intervention**, highlighting how various approaches contribute to improving spatial capability. Additionally, we conduct experiments on 9 benchmarks to assess the spatial capabilities of VLMs. Integrating diverse perspectives, this survey establishes a coherent structure and guiding framework to future research in this rapidly evolving area. Specifically, our key contributions are summarized as follows:

- **An Intervention-Based Taxonomy:** We introduce a five-category taxonomy that systematizes current research by the level of intervention. The categories include: 1) training-free prompting, 2) model-centric enhancements, 3) explicit 2D information injection, 4) 3D spatial enrichment, and 5) data-centric approaches.
- **A Systematic Review of the Ecosystem:** We provide a holistic analysis of the field, investigating the root causes of spatial reasoning failures and comprehensively mapping the landscape of applications, specialized tasks, datasets, and evaluation protocols.
- **A Large-Scale Empirical Evaluation.** We conduct rigorous experiments on 9 widely-used benchmarks to assess the true capabilities of modern VLMs. Our evaluation is unprecedented in scale, analyzing 37 distinct models across three crucial classes: 4 commercial systems, 6 leading open-source generalist VLMs, and 27 specialist spatial VLMs.
- **Forward-Looking Analysis and Future Directions.** Our empirical analysis uncovers critical limitations and systemic biases in current benchmarks. Based on these findings, we provide concrete insights and establish a clear agenda with specific directions for future work toward more robust and capable spatial VLMs.

Survey Structure: As shown in Fig. 1, we organize this survey as follows: § 2 introduces the basic concepts of LLMs, VLMs, and the notion of spatial intelligence within these two paradigms. § 3 presents a hierarchical structure of spatial intelligence and summarizes the existing datasets and benchmarks at each level. § 4 provides an in-depth analysis of the potential causes of the weak spatial intelligence observed in current models. In § 5, we review existing methods aimed at improving the spatial capabilities of VLMs. In § 6, we evaluate representative methods on spatial benchmarks. § 7 highlights real-world applications that demonstrate the role of spatial intelligence in VLM-based systems. § 8 discusses the current challenge faced by the spatial AI system and outlines future research directions, followed by concluding remarks in § 9.

2 Background of Spatial VLMs

In this section, we introduce the foundations of spatial intelligence in the context of VLMs. We begin with an overview of human spatial cognition, then summarize the development of VLMs. Subsequently, we define spatial intelligence for VLMs and trace the evolution of spatial understanding in AI, culminating in the rationale for this survey.

2.1 Human Spatial Cognition

Human spatial cognition refers to the set of mental processes involved in perceiving, understanding, remembering, and reasoning about the spatial dimensions of the environment. Recognized as a distinct cognitive element [25, 26], it encompasses a suite of abilities [27], including:

1) **Spatial Perception**: The ability to perceive and visually understand spatial information both within and beyond the environment, encompassing features, properties, measurement, shapes, positions, and motion [28–30].
2) **Mental Rotation and Transformation**: The capability to mentally manipulate objects in 2D or 3D [31–33].
3) **Spatial Memory**: Encoding, storing, and retrieving information about spatial configurations and layouts, forming the basis of cognitive maps [34].
4) **Spatial Reasoning**: The mental ability to understand, manipulate, and navigate objects and their spatial relationships within the physical world [35, 36].

These abilities support a broad spectrum of everyday activities, from reaching for an object to navigating unfamiliar environments. Research in cognitive psychology, neuroscience, and geoinformation science underscores the central role of spatial cognition for human [35, 25, 37]. Understanding these processes provides a foundation for developing analogous capabilities in AI systems. This survey examines recent progress toward this goal within recent VLMs.

2.2 Spatial Intelligence in Large Foundation Models

The impressive reasoning capabilities of Large Language Models (LLMs) (*e.g.,* GPT-4 [38]) have prompted extensive investigation into their capacity for spatial reasoning using text alone. Many studies have demonstrated that LLMs can derive spatial understanding ability from statistical patterns in text, enabling them to represent spatial information [39, 40], and execute simple spatial reasoning tasks [41–43]. However, this linguistic representation is inherently ungrounded[44–46]; it lacks any connection to physical or visual reality [47, 48]. This can lead to plausible but factually incorrect or physically impossible spatial inferences, highlighting a fundamental limitation of text-only models.

Consequently, the critical next step is to ground language in vision. Early multimodality studies [49] attempt this by linking language modules with object detectors, where spatial reasoning is performed via simple heuristics on object coordinates (*e.g.,* comparing the y-values of two bounding boxes to determine which is "above"). Although initially promising, it remains brittle and fails to capture the nuances of human spatial language, motivating more integrated foundation models.

Recent advances in VLMs aim to combine strong visual and linguistic capabilities for spatial reasoning. Despite progress, current VLMs still struggle with fine-grained spatial configurations, which are critical for understanding dynamic interactions and contextual relationships in the physical world [50, 51]. This challenge highlights that true multimodal intelligence requires not only object and semantic recognition but also spatial reasoning, enabling a shift from static perception to situated understanding.

2.3 Rationale for This Survey

As VLMs become increasingly capable of perceiving complex cues on top of the foundational reasoning abilities of LLMs, the challenge of acquiring visual spatial intelligence becomes more prominent. This raises key challenges:

1. *How effectively can VLMs perceive spatial cues from raw visual input?*
2. *How well can they understand spatial relationships among objects in the visual cue?*
3. *How accurately can they align such spatial understanding across modalities to better support spatial reasoning?*

These three factors are interrelated and collectively influence the performance of VLMs on spatial-related tasks. However, the spatial intelligence of VLMs remains an underexplored aspect, posing a significant challenge for the research community. From this perspective, we present this survey to systematically review spatial intelligence of VLMs, aiming to shed light on this latent capability and encourage deeper investigation into its underlying mechanisms.

Spatial reasoning is not merely an academic goal but a prerequisite for real-world AI systems that interact with the physical environment. Strengthening spatial intelligence is therefore central to applications such as autonomous systems [52], robotics [4], augmented reality [53], human-centered decision-making [54], and human–AI collaboration. This survey aims to provide a comprehensive overview of the state of the art, thereby serving as a valuable resource for researchers and practitioners working towards this goal.

3 Spatial Tasks, Datasets, and Benchmarks

To bring structure to the often-conflated field of spatial reasoning, this section systematically organizes the core tasks and the datasets used to train and evaluate them. We first establish a three-level cognitive hierarchy to categorize spatial tasks based on the skills they require (§ 3.1). Following this structure, we then survey the key datasets and benchmarks designed to train and measure performance on these tasks, summarizing their core characteristics (§ 3.2).

3.1 A Cognitive Hierarchy of Spatial Tasks

To move beyond the vague label of "spatial reasoning", we introduce a three-level cognitive hierarchy inspired by *human spatial cognition*: **Perception**, **Understanding**, and **Extrapolation**. This cumulative structure provides a precise vocabulary for analyzing VLM capabilities. It distinguishes between the foundational ability to perceive individual objects (Perception), the more complex skill of reasoning about their relationships (Understanding), and the advanced capability to extrapolate or infer unobserved spatial states (Extrapolation). This framework allows us to systematically map tasks to the specific skills they test, enabling a clearer diagnosis of model failures.

3.1.1 Level 1: Spatial Perception

Spatial Perception encompasses the ability to perceive individual objects along with their intrinsic spatial attributes (*e.g.*, size, geometric structure, and orientation). This capability extends beyond conventional semantic recognition by requiring explicit awareness of spatial properties rather than merely identifying object categories or abstract features. Representative tasks that probe this capability include: (see Fig. 2, upper block).

- **3D Object Detection:** Unlike conventional 2D detection tasks, this task requires spatial awareness across three dimensions [55]. Models must either infer 3D bounding boxes from 2D images [56], or process inputs such as RGB-D [57] and point clouds [58] to capture object dimensions, volume, and geometric structure within 3D space.
- **3D Segmentation:** This task separates individual objects and their closed 3D boundaries in space. The resulting segmentation mask is a direct representation of the object's geometric structure and size [59–61].
- **Orientation Estimation:** This task predicts an object's rotational pose relative to the camera coordinate frame. This fundamental perception skill remains a known weakness even for leading commercial VLMs [38, 62].
- **Depth Estimation:** It evaluates a model's ability to perceive depth. In VLMs, textual information can be integrated to guide depth prediction in the visual branch [63–65]. It also appears as a vision-centric task in VQA settings [66], where models answer depth-related questions.

3.1.2 Level 2: Spatial Understanding

Building on perception, **Spatial Understanding** involves reasoning about the relationships among multiple objects within a scene. This requires interpreting spatial cues such as prepositions, relative directions, and the broader geometric and semantic composition of the environment. At this level, the focus shifts from localizing individual entities to comprehending the overall scene structure, addressing the question: "*How are objects arranged relative to one another?*" In the context of VLMs, spatial understanding is typically evaluated through tasks illustrated in middle block of Fig. 2.

- **Spatial Relations VQA:** This task evaluates a model's reasoning about spatial relationships among objects [67, 16]. Typical questions test understanding of relative positions and absolute distances, probing fine-grained spatial reasoning beyond single object recognition.
- **Spatial Grounding:** It assesses a model's ability to disambiguate and localize the target object from a spatially-grounded language description. Unlike simple object detection, it requires interpreting explicit spatial relationships (*e.g.*, "the dog on the left") to identify the correct target among similar distractors. This skill is a cornerstone of Level 2, as it tests how well a model moves beyond single-object perception to understanding relative arrangement. The output of the task is typically a segmentation mask or bounding box, and recent work [68–70] reflects a growing interest in this fine-grained spatial understanding.

3.1.3 Level 3: Spatial Extrapolation

Spatial Extrapolation represents the highest level of spatial intelligence, requiring a model to reason beyond the immediately perceptible environment. Built upon both perception and understanding, this skill involves inferring hidden states, predicting future configurations, or adopting alternative viewpoints not explicitly shown in the input. Tasks that test this advanced capability fall into two primary classes: 1) predicting changes in the physical state of the scene; 2) reasoning from a specific, situated viewpoint. While the output formats may differ, we summarize several core tasks capturing the extrapolative capabilities of VLMs, as illustrated in Fig. 2 (the bottom block):

- **Spatial Simulation and Inferring:** This task requires VLMs to mentally project or predict unseen or future spatial states based on the current configuration. It includes several key subtasks: **(1)** *Spatial Manipulation*, where the model must reason about the outcome of an object's movement or physical interaction [16]; **(2)** *Occlusion Reasoning*, which involves extrapolating occluded or hidden elements within a scene (*e.g.*, occlusion-based counting [71]); and **(3)** *Mental Rotation*, where the model identifies the correct corresponding object after a transformation in its orientation [72].
- **Spatial Situated Reasoning:** Moving beyond passive observation, this class of tasks tests a model's capacity to act as a situated agent. It emphasizes context-dependent reasoning, anchored to a specific viewpoint,

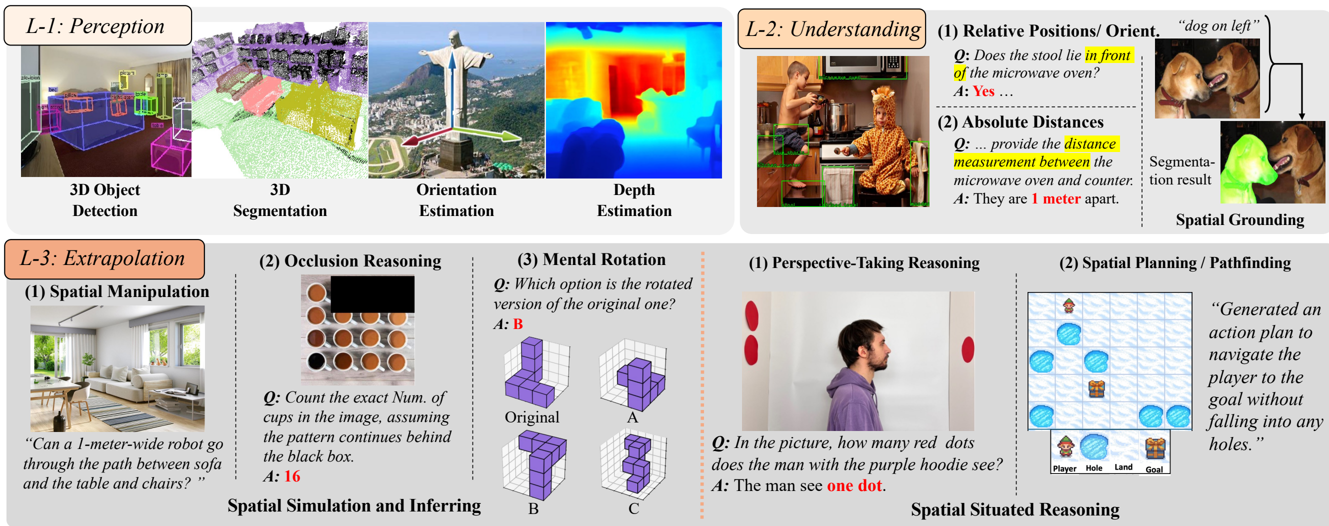

Figure 2 **A three-level hierarchy of spatial intelligence.** *L-1: Perception* involves identifying individual objects and their attributes. *L-2: Understanding* requires reasoning about the relative arrangements between objects. *L-3: Extrapolation* involves inferring unseen states or planning paths.

position, or goal within the scene. Concrete examples include **(1)** *Perspective-Taking Reasoning*, where the model must adopt a designated visual perspective to answer questions accurately [73, 74], and **(2)** *Spatial Planning*, which requires generating a navigation path within a layout (*e.g.*, a map) between given start and goal locations [75, 76].

3.2 Spatially-Oriented Dataset and Benchmark

With our cognitive hierarchy established, we now survey the key datasets and benchmarks that drive progress in spatial AI. To provide a focused analysis, we scope this review to the **Visual Question Answering** (VQA) paradigm, the *de facto* standard for training and evaluating nuanced spatial skills. Our survey concentrates on resources introduced over the past two years that are explicitly designed to assess spatial intelligence. This excludes the resources from general-purpose datasets and benchmarks where spatial reasoning is an implicit or inseparable component [77, 78].

3.2.1 Training Corpora: Growth and Gaps

A recent research effort has begun to address the longstanding scarcity of spatially-oriented training data, evidenced by a notable acceleration in data volume from 2023 to 2025. Our survey identifies **21 dedicated training corpora** from this period, which we organize chronologically in Fig. 3 to illustrate this trend. (Detailed characteristics and corresponding links for each resource are provided in Appendix Tab. 2). Despite this promising progress, our analysis reveals three critical limitations that continue to hinder the development of robust spatial intelligence.

- **1) Cognitive Imbalance.** As shown by the vertical distribution in Fig. 3, current training data is biased toward lower-level cognitive skills. Datasets targeting *perception* and *relational understanding* are more numerous and larger in scale than those designed for *extrapolation*. Besides, key abilities of *extrapolation*, such as mental rotation, are often underrepresented or absent, highlighting a major gap in the resources available for future research.
- **2) Insufficient Scale.** While growing, the largest spatially-dedicated corpora contain millions of examples. This is orders of magnitude smaller than the *billion-scale datasets* used to train foundation models like GPT-4o [38]. This scale disparity limits the development of fundamental spatial capabilities from pre-training alone, forcing reliance on downstream fine-tuning.
- **3) Predominance of Single-View 2D Data.** The vast majority of public training data are built from single-view 2D images. Datasets based on explicit spatial representations, such as multi-view imagery or native 3D data (*e.g.*, point clouds), remain scarce. This modality bias is a critical bottleneck, as constructing and

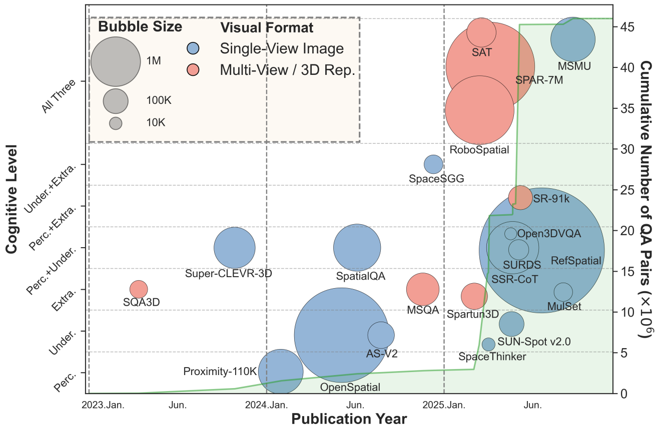

Figure 3 Overview of spatially-oriented training datasets. The figure plots 21 training corpora by their release date and targeted cognitive skill. Bubble size denotes data scale, while bubble color indicates modality (*i.e.*, single-view only, or including multi-view or 3D representation). The cumulative data volume (green line) shows a promising acceleration. (More dataset details are provided in Appendix Tab. 2)

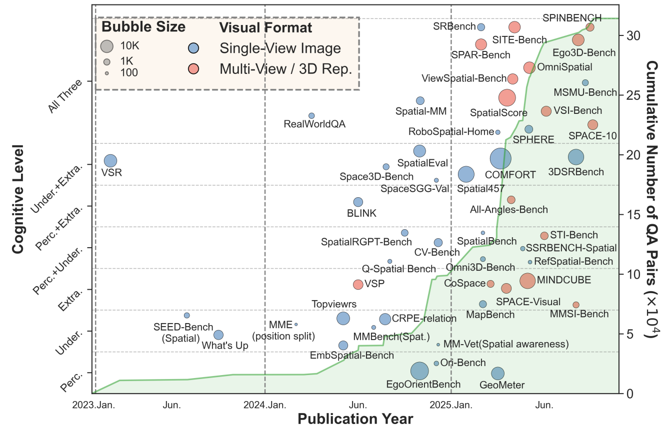

Figure 4 Overview of spatial evaluation benchmarks. This figure maps 49 benchmarks using the same visualization scheme as the training data overview. Compared to the training corpora, the evaluation landscape is notably broader, with a greater number of benchmarks covering a wider range of cognitive skills. (Details for each benchmark are provided in Appendix Tab. 3)

annotating these richer data formats is significantly more challenging, yet they are essential for teaching models about true 3D geometry and overcoming the ambiguities inherent in 2D projections.

3.2.2 Evaluation Benchmarks: A Rapid Expansion

Compared to training datasets, evaluation benchmarks have grown more rapidly. We identify **49 distinct benchmarks** released within the past two years, spanning different cognitive levels (see Fig. 4). More benchmark details are summarized in Appendix Tab. 3.

A key characteristic of these benchmarks is their relatively small scale, typically under 10K samples. However, this smaller scale allows for more meticulous, expert-driven annotation and targeted data curation, resulting in benchmarks that cover more diverse and challenging scenarios than their training-focused counterparts. We also observe an emerging trend of comprehensive benchmarks designed to probe all three cognitive levels within a single evaluation suite. The rapid growth of these diverse, comprehensive diagnostic tools is important in systematically evaluating and revealing the spatial weaknesses of current VLMs.

4 Causes of Deficient Spatial Understanding in General VLMs

Despite remarkable advances, current VLMs exhibit systematic limitations when confronted with tasks requiring precise spatial understanding. As these models are increasingly deployed in applications demanding fine-grained spatial reasoning such as embodied decision making and autonomous navigation, their spatial intelligence deficits become more pronounced and consequential. Extensive empirical studies have documented these limitations across diverse spatial tasks [79–82, 78, 83].

In this section, we review recent research findings to identify the key factors underlying spatial intelligence limitations in VLMs. By examining causes ranging from architectural design to training paradigms, we provide a systematic analysis of why current models struggle with spatial perception, understanding, and extrapolation, and offer insights to guide future improvements.

4.1 Reason I: Suboptimal Vision Encoder

The vision encoder in VLMs provides crucial visual representations for downstream cross-modal reasoning. CLIP [84] has become the de facto choice across open-source VLMs due to its strong generalization and semantically rich embeddings that integrate well with LLMs. However, CLIP exhibits clear limitations in spatially grounded tasks [85–89]. Empirical studies show that it prioritizes global semantics over fine-grained spatial details [87, 86].

Wang *et al.* [90] compared multiple pretrained visual tokenizers, finding that although CLIP excels at semantic alignment, alternative encoders achieve superior performance on fine-grained spatial perception.

Despite these limitations, CLIP-based encoders remain the default choice in most open-source VLMs, as evidenced by BRAVE [91] and OpenVision [92]. Meanwhile, the integration of 3D-aware visual encoders, which is critical for robust spatial comprehension in VLMs [93], remains largely underexplored in current research.

4.2 Reason II: Inefficient Positional Embeddings

ViTs are widely adopted in VLM vision towers, with positional embeddings designed to encode spatial relationships crucial for visuospatial understanding [94]. However, recent research shows that these embeddings are often inefficient or underutilized, limiting models' spatial capabilities.

Two principal failure modes have been identified. First, many positional embedding designs are architecturally suboptimal [95]. Improving the fixed sinusoidal encodings [96] by leveraging a learnable networks or Rotary Position Embeddings (RoPE) [97] have been shown to significantly improve spatial task performance [98, 99]. Second, positional signals are frequently suppressed during computation. Qi *et al.* [50] has shown that high-magnitude vision tokens can "drown out" the subtler positional cues during attention computation, indicating an emergent failure mode where even advanced schemes like RoPE become ineffective under such norm imbalances. These findings suggest the need for norm-aware mechanisms to preserve spatial information. Without sufficient spatial inductive bias, ViTs must learn spatial dependencies purely from data, which is often inefficient and suboptimal [100].

4.3 Reason III: Lack of Effective Modality Alignment

The success of general VLMs lies in effective alignment between modalities. In spatial VLMs, nuanced spatial reasoning demands more fine-grained cross-modal alignment, which is lacking in most current models.

Currently, training paradigms for VLMs focus on semantic alignment at the cost of missing fine-grained details [90]. The dominant reliance on contrastive learning objectives is one of causes of spatial limitation. Although it scales well to web-scale data, this method overlooks fine-grained spatial supervision [101, 84]. Consequently, models tend to optimize for global feature alignment, exhibiting a "bag-of-concepts" behavior [102, 103], which undermines the modeling of spatial relations among objects [78]. Another cause is the lack of fine-grained supervision during training. Dorkenwald *et al.* [98] show that models like Flamingo [104] and GPT-4V [105] underperform on visual localization tasks, due in part to caption-heavy pretraining data with minimal spatial grounding.

Addressing the alignment deficiency, Pandey *et al.* [106] propose a relation-alignment strategy that achieves state-of-the-art performance on Winoground [107]. Wang *et al.* [89] introduce a text-to-pixel alignment method that significantly improves CLIP's performance in referring segmentation, underscoring that global, image-level features alone are insufficient for robust fine-grained spatial reasoning.

4.4 Reason IV: Scarcity of Datasets

Datasets dedicated to spatial understanding across textual and visual modalities remain scarce, posing a significant bottleneck for advancing spatial intelligence in VLMs.

Current datasets used for spatial capability training fall into two categories, each with significant limitations. **General multimodal datasets**, *e.g.,* RefCOCOg [108], GQA [109], and VisualGenome [110] are commonly adopted for fine-grained visual grounding, yet only a fraction of their samples contain explicit spatial annotations.

Spatial-oriented datasets, *e.g.,* VSR [85], CLEVR [111], CLEVR-Ref+ [112], CLEVR-Humans [113] and What's Up [51], are designed to probe spatial reasoning. However, they remain limited in several respects. First, these datasets typically lack annotations for absolute spatial concepts (*e.g.,* precise quantitative distance) and detailed geometric properties (*e.g.,* relative object sizes). Second, their relational diversity is restricted: studies show that the top 17% of relation categories account for over 90% of all spatial examples [114, 115], resulting in an bias toward common spatial relations. Third, they fail to support spatial extrapolation tasks, which require higher-order reasoning beyond direct observation. Finally, their scale and diversity remain far below the web-scale datasets that have fueled semantic-level pretraining. Due to the scarcity of public spatial datasets, recent studies [67, 116–118, 82, 119–123, 93] have to construct their own spatially annotated datasets to advance research.

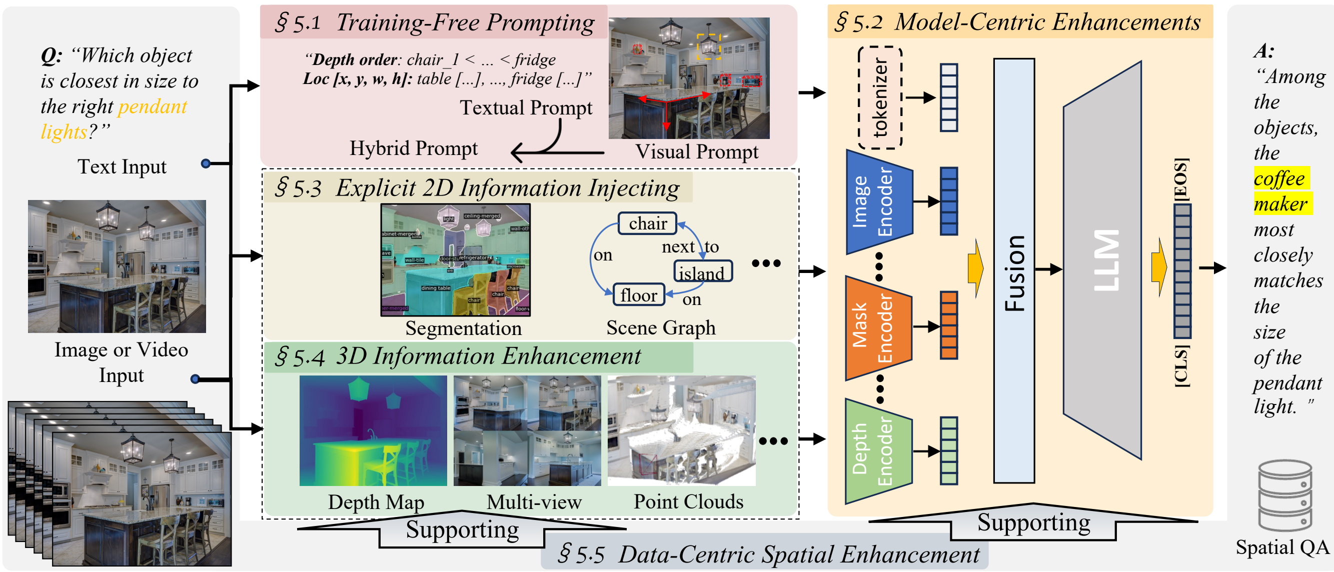

Figure 5 A schematic overview linking methods from § 5.1–5.5 to corresponding modules in the VLM framework.

5 Methods for VLM Spatial Enhancement

Recognizing the limitations of spatial intelligence in current VLMs, the research community has explored different strategies to enhance their spatial capabilities. In this section, we introduce an **intervention-based taxonomy** that systematically organizes recent approaches according to the level of intervention applied to improve spatial awareness. Specifically, we categorize methods into key directions: **Training-Free Prompting** (§ 5.1), **Model-Centric Enhancement** (§ 5.2), **Spatial Feature Augmentation** (§ 5.3, § 5.4), and **Data-Centric Spatial Enhancement** (§ 5.5). This taxonomy provides a structured view of how different methodological interventions target complementary aspects of the spatial intelligence challenge.

5.1 Training-Free Prompting Methods

Recent studies demonstrate that prompt engineering can markedly enhance model reasoning performance [124]. In the context of spatial intelligence, prompting offer a training-free pathway to improve VLM spatial reasoning performance at inference. We categorize existing methods by input space into three groups: (1) textual prompting, (2) visual prompting, and (3) hybrid prompting.

5.1.1 Textual Prompting Methods

Textual prompting methods enhance spatial reasoning by explicitly highlighting spatial cues, object relationships, and orientations within language instructions.

One prominent approach leverages Chain-of-Thought reasoning [125] to decompose spatial tasks into intermediate steps. For example, SpatialPrompt [12] first instructs the VLM to identify reference objects, and then reasoning based on such references. Similarly, Mitra *et al.* [126] prompt VLMs to generate scene graphs that capture spatial relationships, which are subsequently used alongside the original image and task description to enable a better inference process.

Another stream of research leverages expert models to extract detailed spatial information, as the enhanced textual priors for spatial reasoning. Zhao *et al.* [127] use expert models to extract 2D positional information and inter-entity relationships, which are then integrated into textual prompts to guide the reasoning. SpatialPin [128] leverages a language-guided segmentation model with camera pose estimation, depth prediction, and 3D scene reconstruction to generate fine-grained spatial descriptions that enrich the textual prompt. Similarly, SOFAR [129] integrates detection, segmentation, and orientation estimation models to construct 6-DoF scene graphs, enabling enhanced spatial planning in zero-shot scenarios.

5.1.2 Visual Prompting Methods

Visual prompting uses additional visual cues or auxiliary visual inputs to enhance models' scene understanding.

One effective strategy is to incorporate additional visual features directly into the prompt. For instance, applying masks over instances can suppress attention to weakly related regions while maintaining spatial coherence between targets and their surrounding context [130]. Such a method has been shown to outperform fully fine-tuned state-of-the-art models on referring expression comprehension in zero-shot settings [131]. Building on this idea, 3DAxisPrompt [132] enhances spatial reasoning by superimposing a 3D coordinate system onto images, further strengthening the model's spatial awareness.

Another direction leverages richer spatial information from multiple images or videos. Initializing from the single image input, ZeroVLM [133] uses Zero-1-to-3 [134] to synthesize novel views and concatenate them to be fed into VLM for question answering. MindJourney [135] uses video diffusion to generate trajectories within a scene, enabling the VLM to reason over multi-view evidence accumulated through interactive exploration. Tackling video data, Coarse Correspondences Prompting [136] tracks objects, and extracts coarse correspondences in key frames as visual prompts, helping VLMs capture spatial–temporal dynamics more effectively.

5.1.3 Hybrid Prompting

Hybrid prompting approaches integrate spatial information across modalities during inference, combining both visual and textual cues to enhance spatial understanding. Lei *et al.* [137] propose Scaffolding Prompting, which overlays a labeled dot matrix onto the image while incorporating the corresponding coordinates into the text prompt, thereby improving vision–language alignment. Alternatively, SeeGround [138] enhances spatial understanding by combining spatially enriched textual descriptions derived from 3D scenes with query-aligned labeled images. Similarly, SpatialPrompting [139] selects 2D keyframes aligned with 3D embeddings, and the original textual input is augmented (by *e.g.,* camera position) together with the keyframes to enrich spatial detail.

Another line of research draws inspiration from human reasoning by integrating multimodal information into the reasoning process. Lee *et al.* [140] use a numerical textual prompt or an abstract visual prompt to enhance perspective-aware reasoning tasks. Simulating mental visualization, Wu *et al.* [43] enhance spatial reasoning by visualizing intermediate chain-of-thought steps and conditioning the final inference on both visual and textual reasoning signals. More advanced methods, Zhou *et al.* [141] and Hu *et al.* [142], extend this idea by enabling VLMs to invoke external tools during inference. In these approaches, the model automatically selects tools to further process visual inputs in conjunction with textual rationales. Like using "sketches" in human reasoning, this strategy substantially improves spatial reasoning performance.

5.2 Model-Centric Enhancements

Utilization of visual spatial information is the key for strong spatial intelligence. However, VLMs still struggle to capture such fine-grained visual details (§ 4.1, § 4.2, and § 4.3). These models often over-rely on language-based reasoning [143], which fails to compensate for their weak spatial grounding.

In this section, we categorize existing approaches to enhancing spatial intelligence into three key perspectives: (1) advanced training and data strategies, (2) architectural enhancements, and (3) encoder-level improvements.

5.2.1 Advanced Training Strategies

Spatial details within images are often overlooked when VLMs mainly focus on global semantics [78, 50]. To address this, recent work has customized training designs to better capture fine-grained visual–spatial information.

From the perspective of task design during training, ROSS [144] introduces image reconstruction loss, as an auxiliary task, to enhance the ability of capturing fine-grained spatial details. Simulating human's multi-step reasoning, Spatial-CoT [145] introduces spatial coordinate alignment and trains models to first generate a spatial rationale in text, which conditions the subsequent autoregressive task. Cube-LLM [146] train models to first identify 2D objects in vision and then use this information to facilitate 3D object grounding. Mimicking human mental visualization, MVoT [147] allow models to "draw" intermediate visualizations through multistep training, using these to support spatial reasoning.

For training strategy, reinforcement learning has emerged as a promising paradigm. ViLaSR [148] adopts GRPO [149] to acquire mental visualization capabilities. With the same optimization framework, recent works [150–153] have demonstrated notable gains in spatial reasoning. In parallel, other works [154–157] design novel variants of DPO/GRPO, incorporating task-specific policy optimization schemes to boost the spatial reasoning capacity.

Besides reinforcement learning, distillation provides an alternative direction. Concretely, region-level distillation is a straightforward way to learn fine-grained representations. Capturing regional detail, Covert *et al.* [158] trains a student model to reconstruct the teacher's masked output using tokens from the unmasked regions of the input image. Wang *et al.* [90] tune the student's visual tokenizer to align its patch-level features with those generated by EVA-CLIP [159]. LLaVA-AURORA [160] strengthens visuospatial understanding by distilling intrinsic cues such as depth and enforcing the decoding of visual tokens as intermediate steps, guiding the model toward richer spatial perception.

5.2.2 Architectural Enhancements

Beyond training strategies, another line of research attributes limited spatial perception to the model architecture.

Recent works [161, 162] argue that connectors between visual and textual modalities play a critical role in retaining visual spatial cues for spatial tasks. Concretely, Lin *et al.* [161] categorize connectors into feature-preserving connectors and feature-compressing connectors, and find that feature-preserving connectors generally yield better spatial performance. Balancing flexibility and locality preservation, Honeybee [162] proposes the C-Abstractor and D-Abstractor modules to offer a better trade-off between performance and efficiency. Cambrian-1 [13] develops the Spatial Vision Aggregator, an innovative connector that dynamically integrates diverse visual features in support of subsequent LLM reasoning.

The attention mechanism form modern VLMs and manipulating attention maps offers another means of enhancing spatial perception. Chen *et al.* [163] adaptively tune attention weights to enhance spatial reasoning capability. CRG [164] implicitly guide the model's attention by pairing questions with both the original image and a blacked-out counterpart, to emphasize visual–spatial information.

5.2.3 Encoder Improvements

Serving as the "eyes" of a VLM, the vision encoder suffers from training on semantic-level datasets with weak detail alignment, constraining spatial perception. In response, contemporary approaches have prioritized strengthening the granularity of visual feature extraction within vision encoders.

Concretely, SpatialCLIP [165] is designed to replace the original vision encoder in LLaVA, enhancing spatial capability. Alternatively, Yu *et al.* [166] enable the model to decide when to incorporate DINOv2 [167] features to supplement CLIPs' vision embeddings.

Extending this idea, recent works integrate multiple vision encoders to enrich spatial representations. He *et al.* [168] use multitask encoders that combine encoders like VQGAN [169], Pix2Struct [170] and etc. to capture richer visual representations. Poly-Visual-Expert[95] fuses token outputs from multiple experts like SAM[171], LayoutLMv3 [172] and etc. to demonstrate its effectiveness on benchmarks like GQA [173]. SpatialLLM [93] further improves 3D-aware feature learning by combining multiple encoders for more robust 3D spatial reasoning. The benefits of enhanced visual cues for spatial tasks are further evidenced by recent studies [13, 174, 175].

5.3 Explicit 2D Information Injecting

In addition to the approaches discussed earlier, another research stream aims to use the spatial priors extracted from 2D images during model training. Specifically, we categorize these priors into two types: 1) object–region priors and 2) spatial relationship priors.

5.3.1 Object Region Guidance

Given weak spatial reasoning in VLMs, recent work extracts Region of Interest (RoI) and uses region-level information to enhance spatial comprehension of VLMs.

To remedy the weakness in local visual cues, RoI is provided as extra information along with the whole image. Concretely, RegionGPT [176] fuses global and local representations to improve spatial reasoning. Similarly, Chen *et al.* [177] and GPT4ROI [178] employ a region encoder to jointly process specified regions and the full image, offering explicit spatial grounding and enabling fine-grained vision–language alignment. VCoder [179] extends this idea by incorporating segmentation masks as additional image tokens, thereby enriching spatial cues for downstream tasks.

More advanced, Region Selection Token [166] is proposed to automatically identify task-relevant regions. ARGUS [174] grounds goal-directed RoIs into the language generation process. CoVLM [180] inserts communication tokens to propose relevant regions, feeding these features back to the LLM to enhance visual compositional understanding.

Other works exploit localization cues in text. PEVL [181] represents object locations as discrete tokens and jointly modeling them with textual input. Ranasinghe *et al.* [182] incorporate RoI into textual prompts during instruction tuning. He *et al.* [168] further extend text input by incorporating object tags, coordinates, captions and etc., thereby creating structured knowledge that enhances spatial awareness.

Combining textual and visual modalities, Lyrics [183] adapts the querying transformer [184] to align such fine-grained regional priors across modalities.

5.3.2 Scene Graph Guidance

Spatial understanding, beyond instance-level perception, requires comprehension of the relationships among instances. Modeling interrelations among instances is the core challenge in enabling effective spatial understanding. Representing structured spatial knowledge, scene graphs is been used to improve VLMs' spatial intelligence across different levels: Perception, Understanding, and Extrapolation.

Considering the weak compositional understanding of pretrained models, SGVL [185] introduces a scene graph loss to enforce the model's ability to capture fine-grained compositional details. Recently, Assouel *et al.* [186] align the object-centric representations consistent with the corresponding compositional information in scene graphs, thereby improving spatial reasoning performance of VLMs.

Besides the representation learning, Liang *et al.* [187] introduce an interaction-augmented scene graph reasoning framework to enhance the scene understanding of VLMs. To model scenes more explicitly, LLaVA-SG [188] leverages graph neural networks to extract scene graph features in addition to conventional visual and textual embeddings, and feeds these combined representations into the LLM for reasoning. Similarly, Zhao *et al.* [189] employ scene graphs to improve the visual spatial description capability of models.

5.4 3D Spatial Information Enhancement

Training VLMs exclusively on 2D could miss the spatial richness and structural complexity of the 3D physical world. Consequently, advancing VLMs beyond static 2D imagery toward the 3D domain is crucial. This section analyzes this evolution by categorizing enhancement strategies into three primary modalities: (1) explicit 3D geometric representations, (2) implicit 3D information from egocentric views, and (3) hybrid approaches that integrate both data types.

5.4.1 Explicit 3D Geometric Representations

This category of methods leverages direct, explicit 3D data, such as point clouds, voxel grids, and meshes, to provide an unambiguous geometric foundation for spatial reasoning and grounded scene understanding. By grounding VLMs in a precise world coordinate system, these approaches aim to eliminate the ambiguities inherent in 2D projections.

Raw Point Clouds. The most direct approach uses raw point clouds as input, grounding the model in explicit geometric data. These approaches typically follow a common architectural pattern: a pretrained point cloud encoder first extracts scene-level geometric features, which are then tokenized and passed to the VLM. Key innovations lie in how these geometric features are enhanced. In Spatial 3D-LLM [190] and SegPoint [191], they introduce dedicated modules, like progressive spatial awareness scheme or the geometric enhancer with geometry-guided feature propagation module, to refine the extracted features for spatial representations. Alternatively, LL3DA [192] enriches the input before encoding by augmenting the raw point cloud with additional geometric cues like surface normals and height.

Point Clouds with Features. To overcome the semantic sparsity of raw point clouds, other methods enrich point clouds with features lifted from corresponding multi-view images. This augments the sparse geometry with rich semantic and textural information. ScanQA [193] inputs point clouds with lifted image features, using a point cloud encoder and voting module to generate object proposals. A 3D–language fusion layer then models relations among proposals and with question embeddings for downstream reasoning. 3D-LLM [194] renders dense point clouds into multi-view images, lifts extracted features into 3D space, and feeds them into a 3D LLM for downstream tasks.

Voxel Grids. By discretizing a point cloud into a regular 3D grid, these methods can apply powerful 3D CNNs to efficiently extract volumetric features, capturing local geometric patterns in a structured format. 3D-LLaVA [195] clusters features into superpoints, processes them with the Omni Superpoint Transformer, and feeds the resulting tokens into an LLM for multimodal reasoning. SIG3D [196] uses a language-guided estimator to predict an agent's viewpoint, re-encodes features from this perspective, and fuses them with language tokens for reasoning. LSceneLLM [197] selects task-relevant regions via language attention, extracts fine-grained features, and fuses them with coarse features before LLM input.

Depth Maps. Depth serves as a key aspect of spatial information, and recent studies have incorporated it into VLMs. VCoder [179] introduces a versatile vision encoder that incorporates depth maps as additional inputs to enhance the perception capability. Similarly, in SpatialBot [15], SpatialRGPT [67], and SmolRGPT [198], they customized depth modules are introduced to handle spatial information derived from a frozen visual backbone. RoboRefer [17] trains a depth encoder and its projection module alongside a conditional RGB encoder to strengthen the VLM's spatial representations. In parallel, SSR [199] leverages intermediate latent rationale tokens derived from depth maps to guide response generation. Novelly, SD-VLM [200] introduces Depth Positional Encoding, which encodes depth maps into depth-aware positional embeddings, enabling straightforward fusion through element-wise addition.

5.4.2 Implicit 3D from Egocentric Views

Beyond methods requiring explicit 3D scans, other approaches derive spatial understanding from sequences of 2D egocentric inputs (*e.g.,* multiview images, first-person video). These data sources are more available and encode implicit 3D information through perspective, parallax, and motion cues. Models are trained to infer spatial context from situated viewpoints, providing a temporally grounded perspective that is often missing from static 3D scans.

VLM-3R [201] and Spatial-MLLM [202] take multi-view images as input, extracting 2D features using a visual encoder and 3D features using a dedicated spatial encoder (*e.g.,* CUT3R [203] and VGGT [204]). The 2D and 3D features are then fused via a 2D–3D fusion module before being passed into the VLM for multimodal reasoning.

5.4.3 Hybrid Approaches: Fusing Explicit and Implicit Cues

To address unimodal limitations, hybrid approaches have emerged. Explicit 3D data, like point clouds, offers precise geometry but often lacks the rich visual texture and contextual understanding provided by images. Conversely, egocentric views are visually rich but lack explicit, globally consistent geometric structure. Hybrid methods aim to combine these complementary strengths through two main strategies: (1) 3D reconstructing from egocentric images, and (2) jointly modeling of 2D images and 3D representations.

3D Reconstruction from Egocentric Views. These methods operate on 2D image or video inputs, first reconstructing a 3D scene representation and then grounding the VLM in this newly created structure. The primary innovations lie in how 2D features are fused with this reconstructed geometry. For instance, LLaVA-3D [14] and 3D-CLR [205] take multi-view images as input and reconstruct 3D scenes using off-the-shelf methods. After reconstruction, 2D features are extracted from the input images and fused with the 3D representations. LLaVA-3D pools and transforms the fused features into tokens for LLM input, while 3D-CLR applies a 3D–2D alignment loss to distill language-aware 2D features into the 3D voxel grid, resulting in a language-grounded 3D representation for downstream reasoning. Alternatively, SplatTalk [206] leverages multi-view images to train a Gaussian Splatting model, forming spatially rich features that enhance the VLM's spatial reasoning capabilities.

Instead of relying on 2D features, GPT4Scene [207] performs 3D instance segmentation after reconstruction and projects the retrieved objects onto a BEV map. This map, combined with video frames as Spatio-Temporal Object Markers, is fed into a VLM for visual understanding.

Scene-LLM [208] lifts 2D frames and the corresponding features into 3D space to form a unified scene-level representation, enabling models to capture holistic spatial cues.

Joint Modeling of Images and 3D Representations. This dominant strategy assumes both modalities are available as input, fusing 2D images with explicit 3D data (*i.e.,* point clouds) or 2.5D data (*i.e.,* depth maps). These methods can be further categorized by their fusion technique.

A common approach is to first segment the scene into objects, and apply object-centric fusion. In Chat-Scene [209] and Robin3D [210], object-centric 2D and 3D features are extracted based on segmented individual objects. Such features are then projected and tokenized for subsequent reasoning in LLM. Inst3D-LMM [211] follows a similar approach but additionally incorporates depth information. Introducing a Multi-view Cross-Modal Fusion layer and a 3D Instance Spatial Relation module, it integrate 2D and 3D features into relation-aware tokens for enhanced reasoning.

Other methods encode the entire scene globally. DSPNet [212] and LEO [213] employ separate 2D and 3D encoders to extract features from images and 3D inputs, respectively. These features are then either fused through a dedicated fusion module or directly projected into tokens and fed into the VLM for reasoning.

Unique subgroups distinct from the previous ones, ScanReason [214] introduces a unique two-phase paradigm: a visual-centric reasoning module first predicts grounding queries to localize target objects, which are then passed to a subsequent 3D grounding module for spatial reasoning over the original point cloud. MM-Spatial [215] processes multi-view RGB images along with depth maps, incorporating a dedicated depth connector to enhance spatial reasoning.

5.5 Data-Centric Spatial Enhancement

Beyond modifying model itself, a significant line of research focuses on **enhancing the data**. The core principle of this data-centric paradigm is to embed spatial reasoning challenges and their solutions directly into the training corpora. By exposing a model to data rich with explicit spatial relationships, these methods aim to teach fundamental spatial concepts from the ground up. This approach is divided into two primary strategies: (1) augmenting existing real-world datasets and (2) generating novel synthetic data.

5.5.1 Augmenting Datasets with Spatial Annotations

A primary strategy is to enrich existing large-scale datasets with explicit spatial annotations. It is compelling as it leverages the diversity of established corpora (*e.g.,* COCO [216], ScanNet [217]), while avoiding the cost of new data collection. Such enrichment is achieved by adding new annotation layers to 2D images, inferring their 3D structures, or directly annotating 3D scene data.

Enriching 2D datasets, SpaRE [115] leverages a hyper-detailed image–description dataset as input to LLMs to extract spatial reasoning QA pairs. Similarly, AS-V2 [218] employs GPT-4V to construct relation conversation based on COCO images and annotations. By linking model-generated responses to specific regions using location and relation labels, the authors compose 127K annotated dialogues. Pseudo-Q [219] uses region proposals and heuristically derives spatial relationships from relative positions and sizes, pairing them with synthesized training instructions.

Lifting 2D to 3D, SpatialVLM [16, 220] and LLaVA-SpaceSGG[221] adopt this strategy to obtain richer 3D spatial information. Specifically, SpatialVLM[16] uses experts to obtain scene-level captions and extract geometric information from 3D representations. LLaVA-SpaceSGG [221] employs scene-graph captions to capture 2D details and uses a depth estimator to recover 3D structure.

Beyond 2D resources, existing 3D datasets offer another potential for fine-grained spatial annotation. RoboSpatial [19] heuristically mines spatial relationships from point clouds and constructs QA pairs with structured templates. SPARTUN3D [123], using 3RScan [222], generates situated scene graphs and prompts GPT-4o to produce situated captions and QA pairs. Likewise, MSQA [223] obtains various spatial situations by adjusting scene graphs in 3D scenes. Furthermore, MultiSPA [224] and SPAR-7M [18] extract spatial relationships across multiple frames derived from 3D scenes, aiming to enhance multi-frame spatial reasoning.

5.5.2 Synthesizing Data for Spatial Tasks

Beyond existing data resources, synthetic pipelines provide greater flexibility for designing spatial tasks. Sparkle [225] posits that mastering fundamental spatial abilities improves visual–spatial reasoning, and programmatically generates planar node layouts accompanied by Direction, Localization, and Distance QA pairs for training. Wang *et al.* [226] construct orientation-annotated datasets by filtering canonical 3D models from Objaverse [227], labeling object fronts with a 2D VLM, and rendering free-view images annotated with polar, azimuthal, and rotation angles to support orientation estimation. Open3DVQA [228] uses simulators [229] to construct urban environments and forms spatially oriented corpora through templated LLM outputs, after specifying the object and agent angles within each scene.

6 Empirical Evaluation and Analysis

Table 1 Comprehensive comparison of VLMs on spatially oriented benchmarks. The background colors indicate emphasis on different cognitive aspects: *perception*, *understanding*, *extrapolation*, and *all three*. **"–"** indicates that the model is not applicable under the given benchmark. **Bold** is the best and underline is the second best, and * indicates a tailored subset of the original strictly within the target cognition. The description of each dataset provided in the Appendix.

Type / Models / Methods (Datasets)	EgoOrientBench* [230]	GeoMeter* [231]	CV-Bench* [13]	What's Up [51]	SEED-Bench* [232]	SRBench* [233]	MINDCUBE [26]	RealWorldQA [234]	OmniSpatial [235]	Ave.
Commercial Models										
GPT-4o [38]	61.46	65.00	81.92	99.50	69.48	28.90	40.50	63.25	29.09	59.90
GPT-5 [236]	62.54	65.00	91.92	99.63	72.13	**56.20**	42.96	72.03	37.30	**66.63**
Gemini 2.5 flash [237]	56.04	69.00	85.60	99.51	72.72	32.90	37.45	71.11	31.86	61.80
Gemini 2.5 pro [237]	**66.26**	83.00	**92.72**	99.63	73.39	48.10	42.25	73.38	46.86	**69.51**
Open-Source General VLMs										
Qwen2.5-VL-7B [238]	56.44	50.00	78.00	50.92	45.19	29.60	27.59	53.07	31.05	46.87
Qwen2.5-VL-72B [238]	58.94	56.00	87.76	96.72	73.14	42.20	39.94	73.47	46.71	63.88
LLaVA-v1.5-7B [239]	33.72	41.00	62.20	19.02	48.74	28.20	39.11	59.08	35.81	40.76
LLaVA-NeXT-7B [240]	36.77	42.00	63.37	78.17	60.00	27.70	32.87	59.74	38.61	48.80
LLaVA-OneVision-7B [241]	38.67	42.00	65.12	76.32	60.00	33.20	34.47	58.68	39.94	49.82
LLaVA-Next-72B [240]	56.01	**96.00**	88.56	82.07	69.04	40.30	39.75	**73.49**	42.58	65.31
***§ 5.1 Training-Free Prompting Methods** (with Qwen2.5-VL)*										
SpatialPrompt-7B [12]	45.69	47.00	81.84	84.88	64.22	31.70	23.61	60.39	37.18	52.95
SpatialPrompt-72B [12]	65.55	60.00	86.24	93.41	72.17	42.60	39.08	64.58	48.92	63.62
CCOT-7B [126]	46.13	42.00	85.76	92.93	66.73	32.90	24.35	60.78	37.51	54.34
CCOT-72B [126]	51.05	55.00	87.51	96.78	**75.54**	42.90	39.85	73.41	**49.32**	63.48
SoM-7B [131]	56.44	47.00	77.52	72.44	66.61	32.50	27.59	53.07	31.05	51.58
SoM-72B [131]	58.94	56.00	84.72	91.34	69.91	41.60	39.94	73.47	46.71	62.51
Scaffold-7B [137]	43.44	42.00	79.76	74.39	59.45	30.10	23.46	56.73	38.10	49.71
Scaffold-72B [137]	48.04	50.00	87.04	91.59	69.60	34.30	36.49	59.61	45.40	58.01
§ 5.2 Model-Centric Enhancement										
ROSS [144]	35.44	45.00	51.84	8.41	40.92	26.90	34.82	39.08	32.49	34.99
ViLaSR [148]	54.72	50.00	77.84	90.12	65.02	30.10	29.80	61.44	38.81	55.32
M2-Reasoning-7B [151]	44.93	65.00	85.84	92.81	66.97	33.40	35.82	66.27	42.08	59.24
LLaVA-AURORA [160]	35.35	41.00	50.40	**100.00**	37.13	34.70	**44.43**	38.17	29.29	45.61
AdaptVis [163]	33.36	45.00	57.28	24.63	45.99	32.00	43.43	48.89	35.49	40.67
Honeybee [162]	40.80	46.00	36.86	77.44	38.41	24.80	29.43	36.73	33.46	40.44
Cambrian-8B [13]	48.38	32.00	69.04	65.73	66.12	28.50	-	59.87	29.94	49.95
§ 5.3 Explicit 2D Information Injecting										
VPT [166]	33.33	28.00	43.14	63.05	40.06	23.30	33.69	41.83	28.44	37.20
VCoder [179]	38.98	22.00	35.73	14.76	12.23	24.50	1.39	26.80	12.65	21.00
§ 5.4 3D Spatial Information Enhancement										
LLaVA-3D [14]	39.81	42.00	37.20	94.05	34.25	26.20	40.08	40.78	36.93	43.48
SpatialBot-3B [15]	48.64	48.00	65.36	15.24	57.92	36.40	-	57.52	37.51	45.82
VCoder (depth) [179]	39.41	28.00	63.52	43.76	37.92	26.50	2.64	27.36	33.31	33.60
§ 5.5 Data-Centric Spatial Enhancement										
SpaceOm-3B [16]	57.04	43.00	47.56	94.63	62.51	29.20	42.13	58.56	43.70	53.15
SpaceQwen2.5-VL-3B [16]	45.38	28.00	70.00	82.07	63.43	27.60	33.28	46.93	41.88	48.73
SpaceFlorence-2-0.23B [16]	29.23	3.00	16.36	**100.00**	11.62	22.20	-	27.58	15.20	28.15
SpaceThinker-3B [16]	37.60	33.00	66.64	72.93	56.51	28.40	35.67	40.00	31.77	44.72
SpaceMantis-8B [16]	34.49	41.00	61.28	44.63	52.42	33.70	29.03	48.10	36.66	42.37
SpaceLLaVA-13B [16]	32.33	13.00	20.84	85.00	29.66	17.70	24.98	23.66	22.83	30.00
SpaceLLaVA-1.5-7B [16]	41.75	44.00	48.96	28.41	48.99	26.90	38.23	50.07	21.40	38.75

We conduct a large-scale empirical study to assess the true capabilities of modern VLMs and the efficacy of current enhancement methods. We first detail our evaluation settings (§ 6.1) and present the main results in Tab. 1. We then analyze these results through the lens of our cognitive hierarchy (§ 6.3), and evaluate the efficacy of different enhancement methodologies surveyed in § 5 (§ 6.4).

6.1 Experimental Settings

Evaluation Benchmarks. We select a suite of benchmarks based on the principle that they are widely adopted by the research community and emphasize different cognitive levels of spatial understanding. For *Perception*, we use EgoOrientBench [230] and GeoMeter (real world) [231] to test intrinsic attribute understanding. For *Understanding*, we adopt SEED-Bench (spatial relation & instance localization) [232], CV-Bench [13], and What's Up [51] to evaluate relational reasoning. For *Extrapolation*, we use SRBench [72] and MindCube [26] to measure reasoning beyond the immediately perceptible environment. To assess comprehensive capability, we utilize RealWorldQA [234] and OmniSpatial [235], both of which span *All Three* cognitive levels. **Selected Models and Baselines.** Our model selection was designed to cover the full spectrum of available VLMs, with a focus on publicly accessible models for 2D VQA. We group them into three main categories:

- **Commercial VLMs.** We included leading proprietary models (*i.e.,* GPT-4o, GPT-5, Gemini 2.5 flash, Gemini 2.5 pro) to serve as high-performance baselines.
- **Open-Source General VLMs.** We included foundational, non-specialized models (*i.e.,* LLaVA and Qwen2.5) to measure baseline spatial capabilities without fine-tuning.
- **Specialist Spatial VLMs.** We selected representative models from each category surveyed in § 5. Specifically, for prompting-based methods (§ 5.1), we include SpatialPrompt [12] and CCOT [126] for textual prompting; SoM [131] and Scaffold [137] for as visual and hybrid method respectively. All prompting methods are applied to the same Qwen2.5-VL-7B backbone [238]. From § 5.2 model-centric methods and § 5.3 explicit 2D injecting methods, we select ROSS [144], ViLaSR [148], M2-Reasoning [151], LLaVA-AURORA [160], AdaptVis [163], Honeybee [162], Cambrian-8B [13], VPT [166], and VCoder (2D) [179] due to their reproducibility and compatibility with VQA benchmarks. From 3D spatial enhancement methods (§ 5.4), we use LLaVA-3D [14], SpatialBot [15] and VCoder [179] fed with depth information. For § 5.5 data-centric methods, we use the released models trained according to the settings described in corresponding works[235, 16]. To ensure a fair comparison, all selected models are applied for inference across all benchmarks with minimal setting adaptations.

Evaluation Metric. Following standard practice for VQA benchmarks, we use question-answering accuracy as the primary metric for our evaluation.

6.2 Main Experimental Results

The main experimental results are shown in Tab. 1. This table provides a comprehensive comparison of all 37 VLMs (*i.e.,* 4 commercial VLMs, 6 open-source general VLMs, and 28 specialist spatial VLMs), across the 9 selected benchmarks. As shown in Tab. 1, a significant gap in spatial intelligence persists across all three model categories. Even the latest commercial systems fail to achieve consistently high performance, especially on complex extrapolation tasks. In the following section, we will analyze the key trends and insights from this data.

6.3 Performance Across Cognitive Levels

Analyzing performance across different cognitions, we group benchmarks according to their corresponding level. For each cognitive group ,we obtain a single overall average score by averaging the performance of *all* models across *all* benchmarks designated to that specific level. The aggregation results are shown in Fig. 6.

6.3.1 Capability Imbalance across Cognitive Levels

As illustrated in Fig. 6, models excel in *Understanding*-level tasks and most struggle in *Extrapolation*, while the *All Three* setting shows balanced performance that reflects integration across cognitive levels.

Extrapolation requires inference beyond the given visual cues. As illustrated in Fig. 8, a moving-direction question involves first understanding the spatial relationships between frames and then inferring a plausible motion that is not explicitly depicted in the original sequence. Similarly, the mental rotation demands spatial visualization skills to establish correspondences between objects under different orientations. The poor performance in extrapolation

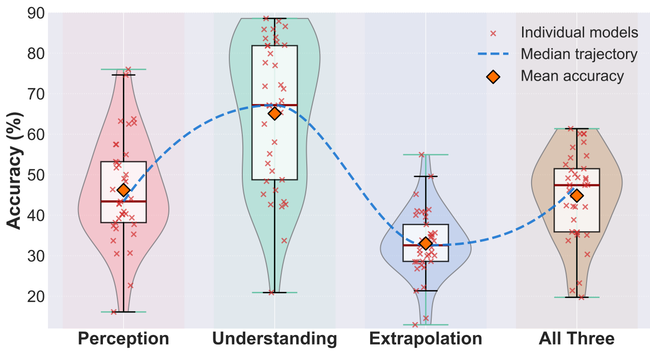

Figure 6 Comparison of model performance across spatial capabilities. Each × denotes the average performance of a method over benchmarks within the same cognitive level.

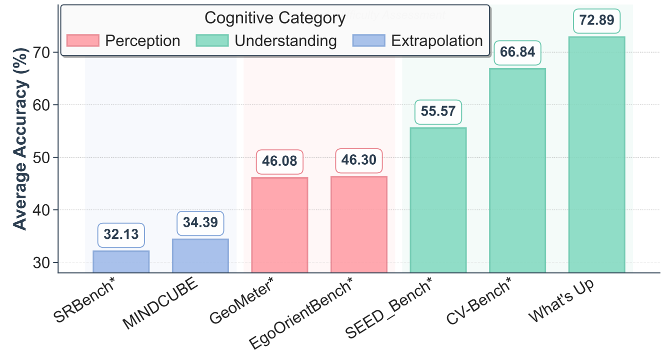

Figure 7 Models' capabilities among different benchmarks. Averaging each model's results on individual datasets, we assess the relative difficulty of each benchmark for VLMs.

tasks reveals a lack of such advanced reasoning capability. In contrast, for *Understanding*-level questions, current VLMs perform better on tasks whose answers can be directly derived from the visual cues shown in Fig. 8.

Surprisingly, perception, which is the foundational capability that supports both understanding and extrapolation in human cognition, exhibits comparatively weak performance in VLMs. This finding highlights the inherent limitation of current models in capturing fine-grained spatial perception.

Key Finding 1: Cognitive Tier Performance Hierarchy

VLMs exhibit a distinct performance hierarchy: *Understanding* (best) > *Perception* > *Extrapolation* (worst). This reveals that current models struggle most with inference beyond observable content, while unexpectedly showing weak fundamental spatial awareness despite their strength in relational reasoning.

6.3.2 Flaws in the Design of Benchmarks

The imbalance across cognitive levels stems not only from model capability but also from the design of existing benchmarks. Reviewing the concrete cases within the datasets in Fig. 8, *Understanding*-level benchmarks such as CV-Bench [13] and What's Up [51] primarily emphasize visual-centric spatial understanding and focus on relative position. However, such designs neglect object-centered cases and other spatial relations between objects, such as relative orientation and absolute distance. In Fig. 7, averaging the models' performance in each individual benchmark, such simplifications in spatial understanding consistently lead to higher performance in the corresponding evaluations.

Key Finding 2: Benchmark Design Bias

Current spatial reasoning benchmarks disproportionately emphasize vision-based relational tasks while underrepresenting metric and object-centric reasoning. This design bias inflates apparent model performance and obscures genuine spatial reasoning limitations.

6.4 Efficacy of Enhancement Methodologies

We now shift our analysis to evaluate the *effectiveness* of different spatial enhancement methods by comparing baseline General VLMs against the various Specialist Spatial VLM categories (Fig. 9). To ensure a fair comparison, this analysis is restricted to models at the 7B parameter scale.

6.4.1 No Universal Solution among Methods

As shown in Fig. 9, no single methodology demonstrates superior performance across all cognitive levels. Instead, we observe a strong method-task alignment: prompting-based strategies clearly outperform others at the *Perception*

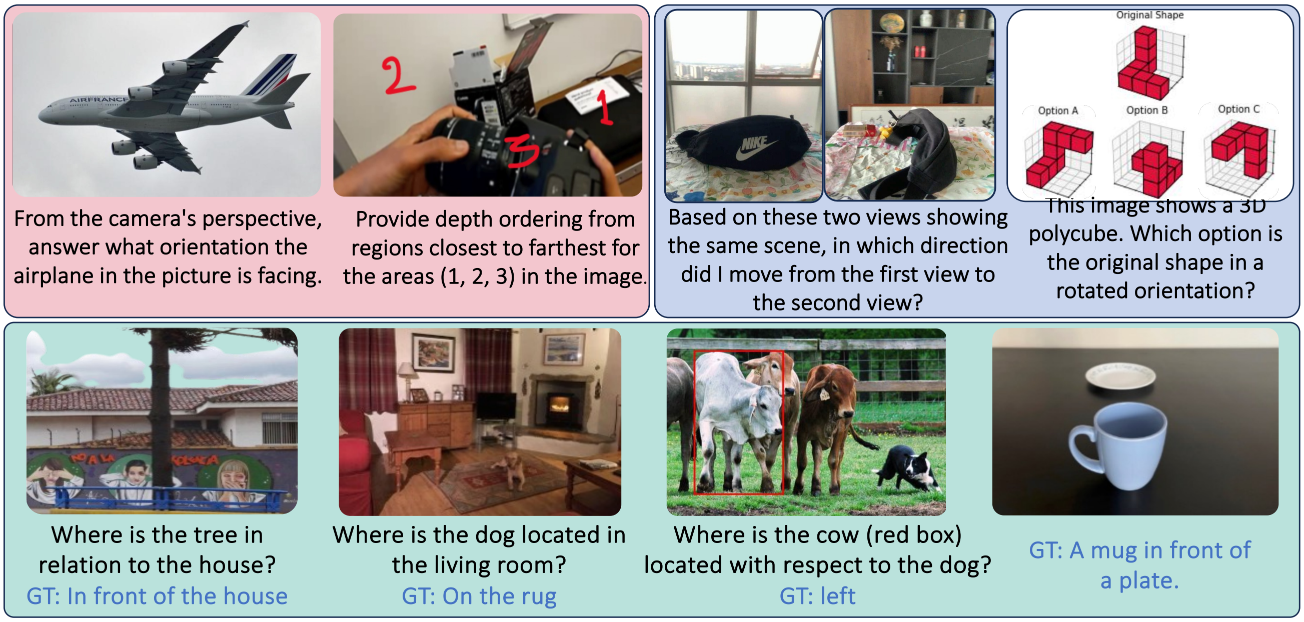

Figure 8 Overview of benchmarks across cognitive levels. Different background colors denote each level: Perception, Understanding and Extrapolation, respectively. The instances are sampled from benchmarks described in Sec. 6.1.

and *Understanding*, whereas data-driven methods perform slightly better in *Extrapolation*-level tasks. In the *All Three* setting, general VLMs show the best overall performance.

Key Finding 3: No Universal Enhancement Strategy

No single spatial enhancement approach dominates across all cognitive tiers. Each method exhibits task-specific advantages: prompting excels at perception/understanding, while data-driven training marginally improves extrapolation. General-purpose models maintain the most balanced cross-tier performance.

6.4.2 Strength and Weakness in Prompting

In Fig. 9, the comparison between general VLM and prompting is both straightforward and intriguing. Modifying the model inputs clearly improves performance in spatial perception and understanding when using the same models on the same questions. However, the prompting strategy has a negative effect on extrapolation tasks. We attribute this weakness to the irrelevance of current spatial prompt designs under the extrapolation setting. Specifically, adding masks[131] and identifying reference objects [12] do not benefit extrapolation, as shown in Fig. 9. Instead, such irrelevant extra prompts introduce additional noise, thereby undermining inference performance.

Key Finding 4: Prompting Trade-offs

Prompt-based interventions demonstrate a clear trade-off: they boost perception and understanding through explicit attention guidance, but degrade extrapolation performance where over-specified cues become distractors rather than aids.

6.4.3 Weak Generalizability in 2D / 3D Enhancements

Unexpectedly, in Fig. 9, methods that explicitly integrate 2D or 3D spatial information exhibit inferior performance under our broad evaluation setting. These results contradict the promising outcomes reported in the original papers and reveal a fundamental issue: the limited generalizability of such models beyond the narrow conditions for which they were originally designed. We therefore recommend that future studies evaluate spatially oriented models

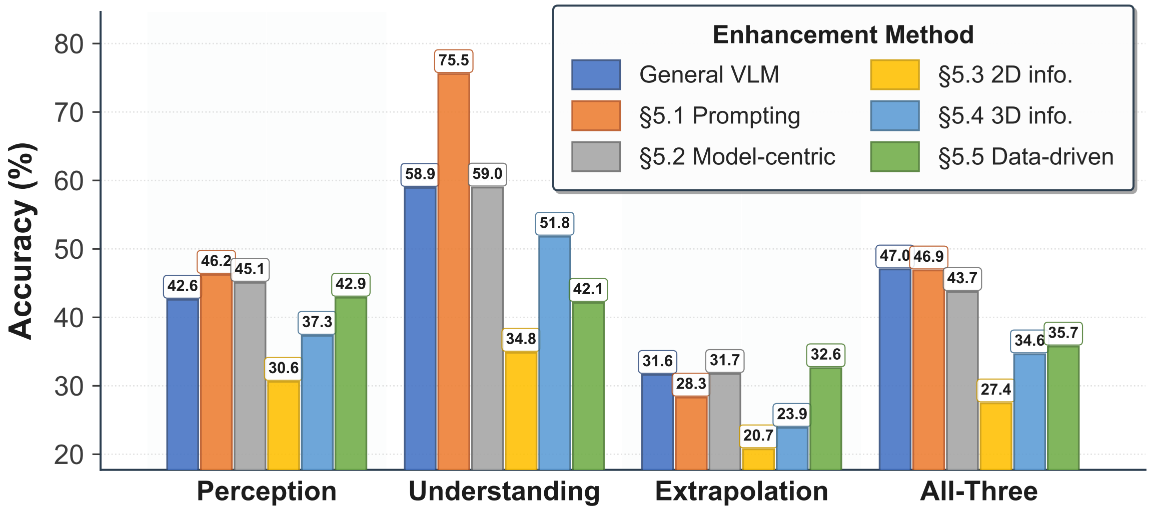

Figure 9 Comparison of performance across different methodologies. Each bin group corresponds to one spatial cognition, and each bar within the group represents the average performance among benchmarks within the same cognition.

under broader and more comprehensive settings that encompass diverse cognitive levels. Spatial intelligence is a broad and multifaceted concept, and our results show that general-purpose VLMs tend to outperform their spatially customized counterparts in comprehensive evaluations.

Key Finding 5: Limited Generalization of Specialized Methods

Explicit 2D/3D spatial enhancements demonstrate limited generalization, with general-purpose models often outperforming specialized variants, revealing the need for more holistic approaches.

7 Spatial VLM Applications

Spatial VLMs are driving major advances across applications ranging from embodied AI systems to AR/VR and next-generation image generation tools. In these domains, spatial intelligence is essential for AI systems to understand 3D environments and follow human instructions more effectively. In this section, we highlight specific applications that can benefit from the enhanced spatial intelligence of VLMs.

Robotics & Embodied AI. Robotic agents process both visual and textual inputs to interact with 3D environments. Within such systems, the spatial capability plays a crucial role in determining downstream performance [242]. Recent studies attribute the limitations of embodied AI to insufficient 3D spatial understanding, and seek to improve it by training with richer 3D representations [243–247]. To improve planning capability, another line of work incorporates graph-based scene representations to capture the spatial configurations [248–252]. Collectively, these studies emphasize that the spatial capability of VLMs is central to the functioning of embodied agents.

Autonomous Driving. Inspired by the capabilities of VLMs, recent works have explored their application in autonomous driving [253, 254], enabling reasoning over dynamic and spatially complex environments. However, the ability of VLMs to comprehend complex spatial relationships remains limited [52, 255], and recent studies [256, 257, 6] demonstrate that enhancing the spatial understanding of VLMs is crucial for tackling high-stakes, real-world driving tasks. Specifically, Zeng *et al.* [257] propose a visual Chain-of-Thought to resolve ambiguities in symbolic planning, and Tian *et al.* [6] integrate 3D-aware scene graphs to enhance VLM reasoning. Complementing these, MPDrive [256] and Reason2Drive [258] employ object detection modules to strengthen spatial grounding and improve scene understanding in complex driving environments.

Image/Video Editing & Generations. Image and video generating and editing require the generation to be aligned with textual instructions while preserving spatial structure[9, 259, 260]. However, research shows that diffusion-based models often fail in spatially grounded tasks [261–263]. To address this limitation, attention manipulation techniques are used to preserve spatial layout [264, 265]. Alternatively, Wang *et al.* [266] integrate LLMs with world models to generate temporally consistent driving videos. Yang *et al.* [267] introduce physics-aware pipelines guided by VLM reasoning.

Medical Diagnosis. Medical imaging interprets volumetric scans (*e.g.*, CT, MRI) to produce clinical outputs. However, 2D-trained VLMs struggle to capture 3D context and align volumetric features with medical language. To address this, recent works [268–270] introduce spatially enhanced architectures that explicitly model volumetric structure. By improving spatial understanding across slices, these methods aim to advance VLM performance in real-world diagnosis.

AR/VR. In AR/VR, VLMs falter without egocentric 3D grounding, as deixis becomes ambiguous, physical plausibility weakens, and spatial memory is lost across steps. Duan *et al.* [271] show that commercial VLMs fail on tightly integrated augmentations, revealing shallow spatial reasoning. To address it, VR Mover [272] fuses user speech, gaze, and gestures into world-anchored API calls to reduce ambiguity. Pei *et al.* [273] couple a "cerebral" language agent with a VLM, injecting the spatial memory and action semantics missing in vanilla VLMs. Collectively, these systems move VLMs from passive perception toward spatially grounded, user-aligned assistance in AR/VR environments.

8 Challenges and Future Directions

After reviewing the current progress of spatial VLMs from multiple perspectives, we further analyze the underlying difficulties that distinguish spatial intelligence from general vision-language modeling. We hope that this discussion can spur deeper reflection and inspire future research directions.

1. **Integrating Visual Spatial Priors into VLMs**: Currently, most VLMs are trained on 2D images, limiting the potential to develop robust spatial understanding in 3D space. How to effectively model spatial information, *e.g.*, geometric constraints and depth, is vital for applications like autonomous driving in 3D world. Alternatively, injecting spatial inductive bias into model architectures is another direction. Since ViTs lack the inductive spatial priors of CNNs [274], they often require more data and training [275, 274]. Recent work explores structural modifications to incorporate spatial bias, improving both efficiency and performance [276, 275, 277].

2. **Balancing Holistic and Fine-Grained Semantics**: Prioritizing holistic semantic alignment, VLMs like CLIP show limited spatial capability. Although emphasizing fine-grained spatial information could improve spatial capability, it could happen at the cost of weakening generalization and zero-shot capability. Thus, advancing spatial intelligence entails a fundamental trade-off between maintaining global semantic coherence and achieving detailed spatial understanding [90]. To address the trade-off, hierarchically cascaded AI systems could offer a promising solution: generalist models first capture broad semantic context, followed by specialized modules that refine spatial understanding. This architectural paradigm offers a more efficient and scalable alternative to monolithic models, representing a promising direction for next-generation AI.

3. **The Ambiguous Nature of Referring**: Linguistic ambiguity poses another challenge for spatial VLMs [82]. Spatial tasks involve both explicit and implicit references, which could be ambiguous. Concretely, in "Alex told Jordan that he was standing behind him", the referents of "he" and "him" are unclear. This ambiguity is further compounded by perspective shifts: references may be egocentric (viewer-centered) or allocentric (agent-centered) [140]. Without explicit context, spatial interpretation can easily become misaligned (Fig. 10). To address this challenge, explicitly modeling viewpoint (reference) information [140] for accurate spatial interpretation has become a vital research topic. Furthermore, integrating multi-agent simulations and dialogue-based grounding holds promise for capturing nuanced spatial references and enabling flexible, human-like perspective shifts.

9 Conclusion

Over the past two years, research on spatial intelligence in VLMs has surged, driven by rapid progress in both data resources and methodologies. Notably, 13 of 21 new training datasets (62%) and 28 of 49 benchmarks (57%) were released in just the first nine months of 2025, out of all datasets introduced since Jan. 2023. Methodologically, 57 of the 102 papers (56%) surveyed were published during this same period, underscoring the field's accelerating momentum. Even so, our evaluation shows that VLMs still fall short of human-level spatial reasoning, with

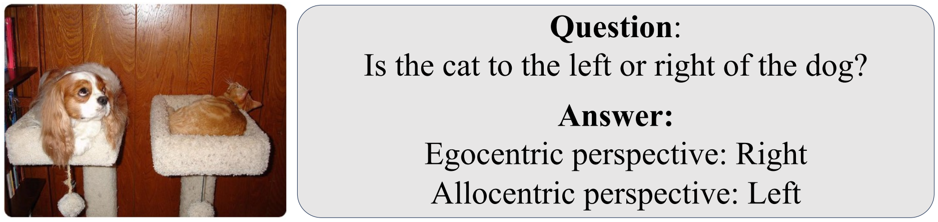

Figure 10 Illustration of referential ambiguity in spatial language. Adopting an *egocentric* (viewer-centered) versus an *allocentric* (agent-centered) perspective leads to two different interpretations of the dog and cat's spatial relationship.

clear imbalances across cognitive levels. These findings highlight the lack of thorough evaluation during model development and the limitations in the current design of spatial data corpora.

References

[1] Jennie E Pyers, Anna Shusterman, Ann Senghas, Elizabeth S Spelke, and Karen Emmorey. Evidence from an emerging sign language reveals that language supports spatial cognition. *Proceedings of the National Academy of Sciences*, 107(27): 12116–12120, 2010.

[2] M Ahmadi and AA Zarei. On the effects of linguistic, verbal, and visual mnemonics on idioms learning. language related research, 12 (5), 279-303, 2021.

[3] Zachariah M Reagh and Michael A Yassa. Object and spatial mnemonic interference differentially engage lateral and medial entorhinal cortex in humans. *Proceedings of the National Academy of Sciences*, 111(40):E4264–E4273, 2014.

[4] Haoyi Zhu, Honghui Yang, Yating Wang, Jiange Yang, Limin Wang, and Tong He. SPA: 3d spatial-awareness enables effective embodied representation. In *The Thirteenth International Conference on Learning Representations*, 2025.

[5] Weizhen Wang, Chenda Duan, Zhenghao Peng, Yuxin Liu, and Bolei Zhou. Embodied scene understanding for vision language models via metavqa, 2025.

[6] Xiaoyu Tian, Junru Gu, Bailin Li, Yicheng Liu, Yang Wang, Zhiyong Zhao, Kun Zhan, Peng Jia, Xianpeng Lang, and Hang Zhao. Drivevlm: The convergence of autonomous driving and large vision-language models, 2024.

[7] Yusi Sun, Haoyan Guan, Yong Hong Kuo, et al. Object-driven narrative in ar: A scenario-metaphor framework with vlm integration. *arXiv preprint arXiv:2504.13119*, 2025.

[8] Ada Yi Zhao, Aditya Gunturu, Ellen Yi-Luen Do, and Ryo Suzuki. Guided reality: Generating visually-enriched ar task guidance with llms and vision models. In *Proceedings of the 38th Annual ACM Symposium on User Interface Software and Technology*, page 1–15. ACM, September 2025. doi: 10.1145/3746059.3747784.

[9] Maxwell J Jacobson and Yexiang Xue. Integrating symbolic reasoning into neural generative models for design generation. *Artificial Intelligence*, 339:104257, 2025.

[10] Jooyeol Yun, Davide Abati, Mohamed Omran, Jaegul Choo, Amirhossein Habibian, and Auke Wiggers. Generative location modeling for spatially aware object insertion, 2024.

[11] Xinhang Liu, Yu-Wing Tai, and Chi-Keung Tang. Agentic 3d scene generation with spatially contextualized vlms. *arXiv preprint arXiv:2505.20129*, 2025.

[12] Yuan-Hong Liao, Rafid Mahmood, Sanja Fidler, and David Acuna. Reasoning paths with reference objects elicit quantitative spatial reasoning in large vision-language models, 2024.

[13] Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai Charitha Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, Ziteng Wang, Rob Fergus, Yann LeCun, and Saining Xie. Cambrian-1: A fully open, vision-centric exploration of multimodal llms, 2024.

[14] Chenming Zhu, Tai Wang, Wenwei Zhang, Jiangmiao Pang, and Xihui Liu. Llava-3d: A simple yet effective pathway to empowering lmms with 3d-awareness. *arXiv preprint arXiv:2409.18125*, 2024.

[15] Wenxiao Cai, Iaroslav Ponomarenko, Jianhao Yuan, Xiaoqi Li, Wankou Yang, Hao Dong, and Bo Zhao. Spatialbot: Precise spatial understanding with vision language models. *arXiv preprint arXiv:2406.13642*, 2024.

[16] 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*, pages 14455–14465, 2024.

[17] Enshen Zhou, Jingkun An, Cheng Chi, Yi Han, Shanyu Rong, Chi Zhang, Pengwei Wang, Zhongyuan Wang, Tiejun Huang, Lu Sheng, et al. Roborefer: Towards spatial referring with reasoning in vision-language models for robotics. *arXiv preprint arXiv:2506.04308*, 2025.

[18] Jiahui Zhang, Yurui Chen, Yanpeng Zhou, Yueming Xu, Ze Huang, Jilin Mei, Junhui Chen, Yu-Jie Yuan, Xinyue Cai, Guowei Huang, et al. From flatland to space: Teaching vision-language models to perceive and reason in 3d. *arXiv preprint arXiv:2503.22976*, 2025.

[19] 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 *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 15768–15780, 2025.

[20] Jingyi Zhang, Jiaxing Huang, Sheng Jin, and Shijian Lu. Vision-language models for vision tasks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[21] Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. A survey on multimodal large language models. *National Science Review*, 11(12), November 2024. ISSN 2053-714X. doi: 10.1093/nsr/nwae403.

[22] Jirong Zha, Yuxuan Fan, Xiao Yang, Chen Gao, and Xinlei Chen. How to enable llm with 3d capacity? a survey of spatial reasoning in llm. *arXiv preprint arXiv:2504.05786*, 2025.

[23] Xianzheng Ma, Yash Bhalgat, Brandon Smart, Shuai Chen, Xinghui Li, Jian Ding, Jindong Gu, Dave Zhenyu Chen, Songyou Peng, Jia-Wang Bian, et al. When llms step into the 3d world: A survey and meta-analysis of 3d tasks via multi-modal large language models. *arXiv preprint arXiv:2405.10255*, 2024.

[24] Jie Feng, Jinwei Zeng, Qingyue Long, Hongyi Chen, Jie Zhao, Yanxin Xi, Zhilun Zhou, Yuan Yuan, Shengyuan Wang, Qingbin Zeng, et al. A survey of large language model-powered spatial intelligence across scales: Advances in embodied agents, smart cities, and earth science. *arXiv preprint arXiv:2504.09848*, 2025.

[25] Marc H Bornstein. Frames of mind: The theory of multiple intelligences, 1986.

[26] Baiqiao Yin, Qineng Wang, Pingyue Zhang, Jianshu Zhang, Kangrui Wang, Zihan Wang, Jieyu Zhang, Keshigeyan Chandrasegaran, Han Liu, Ranjay Krishna, et al. Spatial mental modeling from limited views. *arXiv preprint arXiv:2506.21458*, 2025.

[27] Tyrone Donnon, Jean-Gaston DesCôteaux, and Claudio Violato. Impact of cognitive imaging and sex differences on the development of laparoscopic suturing skills. *Canadian journal of surgery*, 48(5):387, 2005.

[28] Marcia C Linn and Anne C Petersen. Emergence and characterization of sex differences in spatial ability: A meta-analysis. *Child development*, pages 1479–1498, 1985.

[29] Miftakhul Rohmah, Diari Indriati, et al. Hass's theory: how is the students' spatial intelligence in solving problems? In *International conference of mathematics and mathematics education (I-CMME 2021)*, pages 169–175. Atlantis Press, 2021.

[30] Alison Simmons. Spatial perception from a cartesian point of view. *Philosophical Topics*, 31(1/2):395–423, 2003.

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

[32] Stephen Michael Kosslyn. *Image and mind*. Harvard University Press, 1980.

[33] Ronen Porat, Ciprian Ceobanu, et al. Spatial ability: Understanding the past, looking into the future. *European Proceedings of Educational Sciences*, 2023.

[34] Edward C Tolman. Cognitive maps in rats and men. *Psychological review*, 55(4):189, 1948.

[35] John O'keefe and Lynn Nadel. Précis of o'keefe & nadel's the hippocampus as a cognitive map. *Behavioral and Brain Sciences*, 2(4):487–494, 1979.

[36] Mary Hegarty. Mechanical reasoning by mental simulation. *Trends in cognitive sciences*, 8(6):280–285, 2004.

[37] Jean Piaget. *Child's Conception of Space: Selected Works vol 4*. Routledge, 2013.

[38] 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.

[39] Fangjun Li, David C Hogg, and Anthony G Cohn. Advancing spatial reasoning in large language models: An in-depth evaluation and enhancement using the stepgame benchmark. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18500–18507, 2024.

[40] Yutaro Yamada, Yihan Bao, Andrew Kyle Lampinen, Jungo Kasai, and Ilker Yildirim. Evaluating spatial understanding of large language models. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856.

[41] Ishika Singh, Valts Blukis, Arsalan Mousavian, Ankit Goyal, Danfei Xu, Jonathan Tremblay, Dieter Fox, Jesse Thomason, and Animesh Garg. Progprompt: program generation for situated robot task planning using large language models. *Autonomous Robots*, 47(8):999–1012, 2023.

[42] Abhinav Rajvanshi, Karan Sikka, Xiao Lin, Bhoram Lee, Han-Pang Chiu, and Alvaro Velasquez. Saynav: Grounding large language models for dynamic planning to navigation in new environments. In *Proceedings of the International Conference on Automated Planning and Scheduling*, volume 34, pages 464–474, 2024.

[43] 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, 2024.

[44] Yonatan Bisk, Ari Holtzman, Jesse Thomason, Jacob Andreas, Yoshua Bengio, Joyce Chai, Mirella Lapata, Angeliki Lazaridou, Jonathan May, Aleksandr Nisnevich, et al. Experience grounds language. *arXiv preprint arXiv:2004.10151*, 2020.

[45] Roma Patel and Ellie Pavlick. Mapping language models to grounded conceptual spaces. In *International conference on learning representations*, 2022.

[46] Gorka Azkune, Ander Salaberria, and Eneko Agirre. Grounding spatial relations in text-only language models. *Neural Networks*, 170:215–226, February 2024. ISSN 0893-6080. doi: 10.1016/j.neunet.2023.11.031.

[47] 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. *Advances in Neural Information Processing Systems*, 37:75392–75421, 2024.

[48] Xiao Liu, Da Yin, Yansong Feng, and Dongyan Zhao. Things not written in text: Exploring spatial commonsense from visual signals. *arXiv preprint arXiv:2203.08075*, 2022.

[49] Kexin Yi, Jiajun Wu, Chuang Gan, Antonio Torralba, Pushmeet Kohli, and Josh Tenenbaum. Neural-symbolic vqa: Disentangling reasoning from vision and language understanding. *Advances in neural information processing systems*, 31, 2018.

[50] Jianing Qi, Jiawei Liu, Hao Tang, and Zhigang Zhu. Beyond semantics: Rediscovering spatial awareness in vision-language models. *arXiv preprint arXiv:2503.17349*, 2025.

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

[52] Xianda Guo, Ruijun Zhang, Yiqun Duan, Yuhang He, Chenming Zhang, Shuai Liu, and Long Chen. Drivemllm: A benchmark for spatial understanding with multimodal large language models in autonomous driving. *arXiv e-prints*, pages arXiv–2411, 2024.

[53] Uttamasha Monjoree and Wei Yan. Ai's spatial intelligence: Evaluating ai's understanding of spatial transformations in psvt: R and augmented reality. *arXiv preprint arXiv:2411.06269*, 2024.

[54] Zhe Hu, Yixiao Ren, Guanzhong Liu, Jing Li, and Yu Yin. Viva+: Human-centered situational decision-making. *arXiv preprint arXiv:2509.23698*, 2025.

[55] Justin Lazarow, David Griffiths, Gefen Kohavi, Francisco Crespo, and Afshin Dehghan. Cubify anything: Scaling indoor 3d object detection. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 22225–22233, 2025.

[56] Jang Hyun Cho, Boris Ivanovic, Yulong Cao, Edward Schmerling, Yue Wang, Xinshuo Weng, Boyi Li, Yurong You, Philipp Krähenbühl, Yan Wang, et al. Language-image models with 3d understanding. *arXiv preprint arXiv:2405.03685*, 2024.

[57] Jens Piekenbrinck, Alexander Hermans, Narunas Vaskevicius, Timm Linder, and Bastian Leibe. Rgb-d cube r-cnn: 3d object detection with selective modality dropout. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1997–2006, 2024.

[58] Yang Cao, Zeng Yihan, Hang Xu, and Dan Xu. Coda: Collaborative novel box discovery and cross-modal alignment for open-vocabulary 3d object detection. *Advances in Neural Information Processing Systems*, 36:71862–71873, 2023.

[59] Runyu Ding, Jihan Yang, Chuhui Xue, Wenqing Zhang, Song Bai, and Xiaojuan Qi. Pla: Language-driven open-vocabulary 3d scene understanding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7010–7019, 2023.

[60] Songyou Peng, Kyle Genova, Chiyu Jiang, Andrea Tagliasacchi, Marc Pollefeys, Thomas Funkhouser, et al. Openscene: 3d scene understanding with open vocabularies. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 815–824, 2023.

[61] Li Jiang, Shaoshuai Shi, and Bernt Schiele. Open-vocabulary 3d semantic segmentation with foundation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21284–21294, 2024.

[62] 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.

[63] Xueting Hu, Ce Zhang, Yi Zhang, Bowen Hai, Ke Yu, and Zhihai He. Learning to adapt clip for few-shot monocular depth estimation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5594–5603, 2024.

[64] Ziyao Zeng, Daniel Wang, Fengyu Yang, Hyoungseob Park, Stefano Soatto, Dong Lao, and Alex Wong. Wordepth: Variational language prior for monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9708–9719, 2024.

[65] Jinchang Zhang and Guoyu Lu. Vision-language embodiment for monocular depth estimation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 29479–29489, 2025.

[66] Zhipeng Cai, Ching-Feng Yeh, Hu Xu, Zhuang Liu, Gregory Meyer, Xinjie Lei, Changsheng Zhao, Shang-Wen Li, Vikas Chandra, and Yangyang Shi. Depthlm: Metric depth from vision language models. *arXiv preprint arXiv:2509.25413*, 2025.

[67] 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. *arXiv preprint arXiv:2406.01584*, 2024.

[68] Haoxuan You, Haotian Zhang, Zhe Gan, Xianzhi Du, Bowen Zhang, Zirui Wang, Liangliang Cao, Shih-Fu Chang, and Yinfei Yang. Ferret: Refer and ground anything anywhere at any granularity. *arXiv preprint arXiv:2310.07704*, 2023.

[69] Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. Lisa: Reasoning segmentation via large language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9579–9589, 2024.

[70] Hanoona Rasheed, Muhammad Maaz, Sahal Shaji, Abdelrahman Shaker, Salman Khan, Hisham Cholakkal, Rao M Anwer, Eric Xing, Ming-Hsuan Yang, and Fahad S Khan. Glamm: Pixel grounding large multimodal model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13009–13018, 2024.

[71] Atin Pothiraj, Elias Stengel-Eskin, Jaemin Cho, and Mohit Bansal. Capture: Evaluating spatial reasoning in vision language models via occluded object counting. *arXiv preprint arXiv:2504.15485*, 2025.

[72] Ilias Stogiannidis, Steven McDonagh, and Sotirios A Tsaftaris. Mind the gap: Benchmarking spatial reasoning in vision-language models. *arXiv preprint arXiv:2503.19707*, 2025.

[73] Yue Zhang, Zhiyang Xu, Ying Shen, Parisa Kordjamshidi, and Lifu Huang. SPARTUN3d: Situated spatial understanding of 3d world in large language model. In *The Thirteenth International Conference on Learning Representations*, 2025.

[74] 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.

[75] Qiucheng Wu, Handong Zhao, Michael Saxon, Trung Bui, William Yang Wang, Yang Zhang, and Shiyu Chang. Vsp: Assessing the dual challenges of perception and reasoning in spatial planning tasks for vlms. *arXiv preprint arXiv:2407.01863*, 2024.

[76] Gengze Zhou, Yicong Hong, and Qi Wu. Navgpt: Explicit reasoning in vision-and-language navigation with large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 7641–7649, 2024.

[77] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8317–8326, 2019.

[78] Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. When and why vision-language models behave like bags-of-words, and what to do about it? In *International Conference on Learning Representations*, 2023.

[79] Martha Lewis, Nihal V. Nayak, Peilin Yu, Qinan Yu, Jack Merullo, Stephen H. Bach, and Ellie Pavlick. Does clip bind concepts? probing compositionality in large image models, 2024.

[80] Yutaro Yamada, Yingtian Tang, Yoyo Zhang, and Ilker Yildirim. When are lemons purple? the concept association bias of vision-language models, 2024. URL https://arxiv.org/abs/2212.12043.

[81] Tim Brooks, Aleksander Holynski, and Alexei A. Efros. Instructpix2pix: Learning to follow image editing instructions, 2023.

[82] 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, 2025.

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

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

[85] Fangyu Liu, Guy Emerson, and Nigel Collier. Visual spatial reasoning. *Transactions of the Association for Computational Linguistics*, 11:635–651, 06 2023. ISSN 2307-387X. doi: 10.1162/tacl_a_00566.

[86] 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.

[87] Nils Hoehing, Ellen Rushe, and Anthony Ventresque. What's left can't be right – the remaining positional incompetence of contrastive vision-language models, 2023.

[88] Madeline Schiappa, Raiyaan Abdullah, Shehreen Azad, Jared Claypoole, Michael Cogswell, Ajay Divakaran, and Yogesh Rawat. Probing conceptual understanding of large visual-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1797–1807, 2024.

[89] Zhaoqing Wang, Yu Lu, Qiang Li, Xunqiang Tao, Yandong Guo, Mingming Gong, and Tongliang Liu. Cris: Clip-driven referring image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11686–11695, 2022.

[90] Guangzhi Wang, Yixiao Ge, Xiaohan Ding, Mohan Kankanhalli, and Ying Shan. What makes for good visual tokenizers for large language models. *arXiv preprint arXiv:2305.12223*, 2023.

[91] Oğuzhan Fatih Kar, Alessio Tonioni, Petra Poklukar, Achin Kulshrestha, Amir Zamir, and Federico Tombari. Brave: Broadening the visual encoding of vision-language models, 2024.

[92] Xianhang Li, Yanqing Liu, Haoqin Tu, Hongru Zhu, and Cihang Xie. Openvision: A fully-open, cost-effective family of advanced vision encoders for multimodal learning, 2025.

[93] Wufei Ma, Luoxin Ye, Nessa McWeeney, Celso M de Melo, Jieneng Chen, and Alan Yuille. Spatialllm: A compound 3d-informed design towards spatially-intelligent large multimodal models, 2025.

[94] Wanyue Zhang, Yibin Huang, Yangbin Xu, JingJing Huang, Helu Zhi, Shuo Ren, Wang Xu, and Jiajun Zhang. Why do mllms struggle with spatial understanding? a systematic analysis from data to architecture. *arXiv preprint arXiv:2509.02359*, 2025.

[95] Xiaoran Fan, Tao Ji, Shuo Li, Senjie Jin, Sirui Song, Junke Wang, Boyang Hong, Lu Chen, Guodong Zheng, Ming Zhang, et al. Poly-visual-expert vision-language models. In *First Conference on Language Modeling*.

[96] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[97] Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding, 2023.

[98] Michael Dorkenwald, Nimrod Barazani, Cees G. M. Snoek, and Yuki M. Asano. Pin: Positional insert unlocks object localisation abilities in vlms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13548–13558, June 2024.

[99] Siting Li, Pang Wei Koh, and Simon Shaolei Du. On erroneous agreements of CLIP image embeddings, 2025.

[100] Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. *arXiv preprint arXiv:2309.16588*, 2023.

[101] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR, 2021.

[102] Yingtian Tang, Yutaro Yamada, Yoyo Minzhi Zhang, and Ilker Yildirim. When are lemons purple? the concept association bias of vision-language models. In *The 2023 Conference on Empirical Methods in Natural Language Processing*, 2023.

[103] Martha Lewis, Nihal V Nayak, Peilin Yu, Qinan Yu, Jack Merullo, Stephen H Bach, and Ellie Pavlick. Does clip bind concepts? probing compositionality in large image models. *arXiv preprint arXiv:2212.10537*, 2022.

[104] Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Shiori Sagawa, Jenia Jitsev, Simon Kornblith, Pang Wei Koh, Gabriel Ilharco, Mitchell Wortsman, and Ludwig Schmidt. Openflamingo: An open-source framework for training large autoregressive vision-language models, 2023.

[105] OpenAI team. Gpt-4 technical report. 2023.

[106] Rohan Pandey, Rulin Shao, Paul Pu Liang, Ruslan Salakhutdinov, and Louis-Philippe Morency. Cross-modal attention congruence regularization for vision-language relation alignment. *arXiv preprint arXiv:2212.10549*, 2022.

[107] Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. Winoground: Probing vision and language models for visio-linguistic compositionality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5238–5248, 2022.

[108] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C. Berg, and Tamara L. Berg. Modeling context in referring expressions, 2016.

[109] Joshua Ainslie, James Lee-Thorp, Michiel de Jong, Yury Zemlyanskiy, Federico Lebrón, and Sumit Sanghai. Gqa: Training generalized multi-query transformer models from multi-head checkpoints, 2023.

[110] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Fei-Fei Li. Visual genome: Connecting language and vision using crowdsourced dense image annotations, 2016.

[111] Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning, 2016.

[112] Runtao Liu, Chenxi Liu, Yutong Bai, and Alan L Yuille. Clevr-ref+: Diagnosing visual reasoning with referring expressions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4185–4194, 2019.

[113] Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Judy Hoffman, Li Fei-Fei, C. Lawrence Zitnick, and Ross Girshick. Inferring and executing programs for visual reasoning, 2017.

[114] Palaash Agrawal, Haidi Azaman, and Cheston Tan. Stupd: A synthetic dataset for spatial and temporal relation reasoning. *arXiv preprint arXiv:2309.06680*, 2023.

[115] Michael Ogezi and Freda Shi. Spare: Enhancing spatial reasoning in vision-language models with synthetic data, 2025.

[116] Xuefei Sun, Doncey Albin, Cecilia Mauceri, Dusty Woods, and Christoffer Heckman. Spatial-llava: Enhancing large language models with spatial referring expressions for visual understanding, 2025.

[117] Xingrui Wang, Wufei Ma, Tiezheng Zhang, Celso M de Melo, Jieneng Chen, and Alan Yuille. Spatial457: A diagnostic benchmark for 6d spatial reasoning of large multimodal models, 2025.

[118] 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. *arXiv preprint arXiv:2412.14171*, 2024.

[119] Wenyu Zhang, Wei En Ng, Lixin Ma, Yuwen Wang, Junqi Zhao, Allison Koenecke, Boyang Li, and Lu Wang. Sphere: Unveiling spatial blind spots in vision-language models through hierarchical evaluation, 2025.

[120] Yongsen Mao, Junhao Zhong, Chuan Fang, Jia Zheng, Rui Tang, Hao Zhu, Ping Tan, and Zihan Zhou. Spatiallm: Training large language models for structured indoor modeling, 2025.

[121] 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. *arXiv preprint arXiv:2411.06048*, 2024.

[122] Shuo Xing, Zezhou Sun, Shuangyu Xie, Kaiyuan Chen, Yanjia Huang, Yuping Wang, Jiachen Li, Dezhen Song, and Zhengzhong Tu. Can large vision language models read maps like a human?, 2025.

[123] Yue Zhang, Zhiyang Xu, Ying Shen, Parisa Kordjamshidi, and Lifu Huang. Spartun3d: Situated spatial understanding of 3d world in large language models. *arXiv preprint arXiv:2410.03878*, 2024.

[124] Banghao Chen, Zhaofeng Zhang, Nicolas Langrené, and Shengxin Zhu. Unleashing the potential of prompt engineering for large language models. *Patterns*, 6(6):101260, June 2025. ISSN 2666-3899. doi: 10.1016/j.patter.2025.101260.

[125] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35: 24824–24837, 2022.

[126] Chancharik Mitra, Brandon Huang, Trevor Darrell, and Roei Herzig. Compositional chain-of-thought prompting for large multimodal models, 2024.

[127] Yongqiang Zhao, Zhenyu Li, Zhi Jin, Feng Zhang, Haiyan Zhao, Chengfeng Dou, Zhengwei Tao, Xinhai Xu, and Donghong Liu. Enhancing the spatial awareness capability of multi-modal large language model. *arXiv preprint arXiv:2310.20357*, 2023.

[128] 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. *arXiv preprint arXiv:2403.13438*, 2024.

[129] Zekun Qi, Wenyao Zhang, Yufei Ding, Runpei Dong, Xinqiang Yu, Jingwen Li, Lingyun Xu, Baoyu Li, Xialin He, Guofan Fan, Jiazhao Zhang, Jiawei He, Jiayuan Gu, Xin Jin, Kaisheng Ma, Zhizheng Zhang, He Wang, and Li Yi. Sofar: Language-grounded orientation bridges spatial reasoning and object manipulation, 2025.

[130] 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.

[131] 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.

[132] Dingning Liu, Cheng Wang, Peng Gao, Renrui Zhang, Xinzhu Ma, Yuan Meng, and Zhihui Wang. 3daxisprompt: Promoting the 3d grounding and reasoning in gpt-4o. *Neurocomputing*, 637:130072, 2025.

[133] Zaiqiao Meng, Hao Zhou, and Yifang Chen. I know about" up"! enhancing spatial reasoning in visual language models through 3d reconstruction. *arXiv preprint arXiv:2407.14133*, 2024.

[134] Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3d object. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9298–9309, 2023.

[135] Yuncong Yang, Jiageng Liu, Zheyuan Zhang, Siyuan Zhou, Reuben Tan, Jianwei Yang, Yilun Du, and Chuang Gan. Mindjourney: Test-time scaling with world models for spatial reasoning. *arXiv preprint arXiv:2507.12508*, 2025.

[136] Benlin Liu, Yuhao Dong, Yiqin Wang, Zixian Ma, Yansong Tang, Luming Tang, Yongming Rao, Wei-Chiu Ma, and Ranjay Krishna. Coarse correspondences boost spatial-temporal reasoning in multimodal language model, 2024.

[137] 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.

[138] Rong Li, Shijie Li, Lingdong Kong, Xulei Yang, and Junwei Liang. Seeground: See and ground for zero-shot open-vocabulary 3d visual grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

[139] Shun Taguchi, Hideki Deguchi, Takumi Hamazaki, and Hiroyuki Sakai. Spatialprompting: Keyframe-driven zero-shot spatial reasoning with off-the-shelf multimodal large language models. *arXiv preprint arXiv:2505.04911*, 2025.

[140] 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, 2025.

[141] 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, 2024.

[142] 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, 2024.

[143] Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal llms, 2024.

[144] Haochen Wang, Anlin Zheng, Yucheng Zhao, Tiancai Wang, Zheng Ge, Xiangyu Zhang, and Zhaoxiang Zhang. Reconstructive visual instruction tuning. In *The Thirteenth International Conference on Learning Representations*, 2025.

[145] Yuecheng Liu, Dafeng Chi, Shiguang Wu, Zhanguang Zhang, Yaochen Hu, Lingfeng Zhang, Yingxue Zhang, Shuang Wu, Tongtong Cao, Guowei Huang, Helong Huang, Guangjian Tian, Weichao Qiu, Xingyue Quan, Jianye Hao, and Yuzheng Zhuang. Spatialcot: Advancing spatial reasoning through coordinate alignment and chain-of-thought for embodied task planning, 2025.

[146] Jang Hyun Cho, Boris Ivanovic, Yulong Cao, Edward Schmerling, Yue Wang, Xinshuo Weng, Boyi Li, Yurong You, Philipp Krähenbühl, Yan Wang, and Marco Pavone. Language-image models with 3d understanding, 2024.

[147] 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, 2025.

[148] Junfei Wu, Jian Guan, Kaituo Feng, Qiang Liu, Shu Wu, Liang Wang, Wei Wu, and Tieniu Tan. Reinforcing spatial reasoning in vision-language models with interwoven thinking and visual drawing, 2025.

[149] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

[150] Wufei Ma, Yu-Cheng Chou, Qihao Liu, Xingrui Wang, Celso de Melo, Jianwen Xie, and Alan Yuille. Spatialreasoner: Towards explicit and generalizable 3d spatial reasoning, 2025.

[151] Inclusion AI, Fudong Wang, Jiajia Liu, Jingdong Chen, Jun Zhou, Kaixiang Ji, Lixiang Ru, Qingpei Guo, Ruobing Zheng, Tianqi Li, et al. M2-reasoning: Empowering mllms with unified general and spatial reasoning. *arXiv preprint arXiv:2507.08306*, 2025.

[152] Baining Zhao, Ziyou Wang, Jianjie Fang, Chen Gao, Fanhang Man, Jinqiang Cui, Xin Wang, Xinlei Chen, Yong Li, and Wenwu Zhu. Embodied-r: Collaborative framework for activating embodied spatial reasoning in foundation models via reinforcement learning. *arXiv preprint arXiv:2504.12680*, 2025.

[153] Hongxing Li, Dingming Li, Zixuan Wang, Yuchen Yan, Hang Wu, Wenqi Zhang, Yongliang Shen, Weiming Lu, Jun Xiao, and Yueting Zhuang. Spatialladder: Progressive training for spatial reasoning in vision-language models. *arXiv preprint arXiv:2510.08531*, 2025.

[154] Yifan Shen, Yuanzhe Liu, Jingyuan Zhu, Xu Cao, Xiaofeng Zhang, Yixiao He, Wenming Ye, James Matthew Rehg, and Ismini Lourentzou. Fine-grained preference optimization improves spatial reasoning in vlms. *arXiv preprint arXiv:2506.21656*, 2025.

[155] Kun Ouyang, Yuanxin Liu, Haoning Wu, Yi Liu, Hao Zhou, Jie Zhou, Fandong Meng, and Xu Sun. Spacer: Reinforcing mllms in video spatial reasoning. *arXiv preprint arXiv:2504.01805*, 2025.

[156] Peiyao Wang and Haibin Ling. Svqa-r1: Reinforcing spatial reasoning in mllms via view-consistent reward optimization. *arXiv preprint arXiv:2506.01371*, 2025.

[157] Zhenyu Pan and Han Liu. Metaspatial: Reinforcing 3d spatial reasoning in vlms for the metaverse. *arXiv preprint arXiv:2503.18470*, 2025.

[158] Ian Connick Covert, Tony Sun, James Zou, and Tatsunori Hashimoto. Locality alignment improves vision-language models. In *The Thirteenth International Conference on Learning Representations*, 2025.

[159] Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. Eva-clip: Improved training techniques for clip at scale. *arXiv preprint arXiv:2303.15389*, 2023.

[160] Mahtab Bigverdi, Zelun Luo, Cheng-Yu Hsieh, Ethan Shen, Dongping Chen, Linda G. Shapiro, and Ranjay Krishna. Perception tokens enhance visual reasoning in multimodal language models, 2024.

[161] Junyan Lin, Haoran Chen, Dawei Zhu, and Xiaoyu Shen. To preserve or to compress: An in-depth study of connector selection in multimodal large language models. *arXiv preprint arXiv:2410.06765*, 2024.

[162] Junbum Cha, Wooyoung Kang, Jonghwan Mun, and Byungseok Roh. Honeybee: Locality-enhanced projector for multimodal llm. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13817–13827, 2024.

[163] 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. *arXiv preprint arXiv:2503.01773*, 2025.

[164] David Wan, Jaemin Cho, Elias Stengel-Eskin, and Mohit Bansal. Contrastive region guidance: Improving grounding in vision-language models without training. In *European Conference on Computer Vision*, pages 198–215. Springer, 2024.

[165] Zehan Wang, Sashuai Zhou, Shaoxuan He, Haifeng Huang, Lihe Yang, Ziang Zhang, Xize Cheng, Shengpeng Ji, Tao Jin, Hengshuang Zhao, et al. Spatialclip: Learning 3d-aware image representations from spatially discriminative language. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 29656–29666, 2025.

[166] Runpeng Yu, Xinyin Ma, and Xinchao Wang. Introducing visual perception token into multimodal large language model. *arXiv preprint arXiv:2502.17425*, 2025.

[167] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2024.

[168] Xin He, Longhui Wei, Lingxi Xie, and Qi Tian. Incorporating visual experts to resolve the information loss in multimodal large language models. *arXiv preprint arXiv:2401.03105*, 2024.

[169] Patrick Esser, Robin Rombach, and Björn Ommer. Taming transformers for high-resolution image synthesis, 2021.

[170] Kenton Lee, Mandar Joshi, Iulia Turc, Hexiang Hu, Fangyu Liu, Julian Eisenschlos, Urvashi Khandelwal, Peter Shaw, Ming-Wei Chang, and Kristina Toutanova. Pix2struct: Screenshot parsing as pretraining for visual language understanding, 2023.

[171] 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 preprint arXiv:2408.00714*, 2024.

[172] Yupan Huang, Tengchao Lv, Lei Cui, Yutong Lu, and Furu Wei. Layoutlmv3: Pre-training for document ai with unified text and image masking, 2022.

[173] Drew A Hudson and Christopher D Manning. Gqa: a new dataset for compositional question answering over real-world images. *arXiv preprint arXiv:1902.09506*, 3(8):1, 2019.

[174] Yunze Man, De-An Huang, Guilin Liu, Shiwei Sheng, Shilong Liu, Liang-Yan Gui, Jan Kautz, Yu-Xiong Wang, and Zhiding Yu. Argus: Vision-centric reasoning with grounded chain-of-thought, 2025.

[175] Zhiqi Li, Guo Chen, Shilong Liu, Shihao Wang, Vibashan VS, Yishen Ji, Shiyi Lan, Hao Zhang, Yilin Zhao, Subhashree Radhakrishnan, et al. Eagle 2: Building post-training data strategies from scratch for frontier vision-language models. *arXiv preprint arXiv:2501.14818*, 2025.

[176] Qiushan Guo, Shalini De Mello, Hongxu Yin, Wonmin Byeon, Ka Chun Cheung, Yizhou Yu, Ping Luo, and Sifei Liu. Regiongpt: Towards region understanding vision language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13796–13806, 2024.

[177] Chi Chen, Ruoyu Qin, Fuwen Luo, Xiaoyue Mi, Peng Li, Maosong Sun, and Yang Liu. Position-enhanced visual instruction tuning for multimodal large language models. *arXiv preprint arXiv:2308.13437*, 2023.

[178] Shilong Zhang, Peize Sun, Shoufa Chen, Minn Xiao, Wenqi Shao, Wenwei Zhang, Yu Liu, Kai Chen, and Ping Luo. GPT4roi: Instruction tuning large language model on region-of-interest, 2024.

[179] Jitesh Jain, Jianwei Yang, and Humphrey Shi. Vcoder: Versatile vision encoders for multimodal large language models, 2023.

[180] Junyan Li, Delin Chen, Yining Hong, Zhenfang Chen, Peihao Chen, Yikang Shen, and Chuang Gan. Covlm: Composing visual entities and relationships in large language models via communicative decoding. *arXiv preprint arXiv:2311.03354*, 2023.

[181] Yuan Yao, Qianyu Chen, Ao Zhang, Wei Ji, Zhiyuan Liu, Tat-Seng Chua, and Maosong Sun. Pevl: Position-enhanced pre-training and prompt tuning for vision-language models. *arXiv preprint arXiv:2205.11169*, 2022.

[182] 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.

[183] Junyu Lu, Dixiang Zhang, Songxin Zhang, Zejian Xie, Zhuoyang Song, Cong Lin, Jiaxing Zhang, Bingyi Jing, and Pingjian Zhang. Lyrics: Boosting fine-grained language-vision alignment and comprehension via semantic-aware visual objects. *arXiv preprint arXiv:2312.05278*, 2023.

[184] 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 *International conference on machine learning*, pages 19730–19742. PMLR, 2023.

[185] Roei Herzig, Alon Mendelson, Leonid Karlinsky, Assaf Arbelle, Rogerio Feris, Trevor Darrell, and Amir Globerson. Incorporating structured representations into pretrained vision & language models using scene graphs. *arXiv preprint arXiv:2305.06343*, 2023.

[186] Rim Assouel, Pietro Astolfi, Florian Bordes, Michal Drozdzal, and Adriana Romero-Soriano. Object-centric binding in contrastive language-image pretraining, 2025.

[187] Dayong Liang, Changmeng Zheng, Zhiyuan Wen, Yi Cai, Xiao-Yong Wei, and Qing Li. Seeing beyond the scene: Enhancing vision-language models with interactional reasoning, 2025.

[188] Jingyi Wang, Jianzhong Ju, Jian Luan, and Zhidong Deng. Llava-sg: Leveraging scene graphs as visual semantic expression in vision-language models. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2025.

[189] Yu Zhao, Hao Fei, Wei Ji, Jianguo Wei, Meishan Zhang, Min Zhang, and Tat-Seng Chua. Generating visual spatial description via holistic 3d scene understanding. *arXiv e-prints*, pages arXiv–2305, 2023.

[190] Xiaoyan Wang, Zeju Li, Yifan Xu, Jiaxing Qi, Zhifei Yang, Ruifei Ma, Xiangde Liu, and Chao Zhang. Spatial 3d-llm: Progressive spatial awareness for advanced 3d vision-language understanding.

[191] Shuting He, Henghui Ding, Xudong Jiang, and Bihan Wen. Segpoint: Segment any point cloud via large language model. In *European Conference on Computer Vision*, pages 349–367. Springer, 2024.

[192] Sijin Chen, Xin Chen, Chi Zhang, Mingsheng Li, Gang Yu, Hao Fei, Hongyuan Zhu, Jiayuan Fan, and Tao Chen. Ll3da: Visual interactive instruction tuning for omni-3d understanding reasoning and planning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26428–26438, 2024.

[193] Daichi Azuma, Taiki Miyanishi, Shuhei Kurita, and Motoaki Kawanabe. Scanqa: 3d question answering for spatial scene understanding. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19129–19139, 2022.

[194] Yining Hong, Haoyu Zhen, Peihao Chen, Shuhong Zheng, Yilun Du, Zhenfang Chen, and Chuang Gan. 3d-llm: Injecting the 3d world into large language models. *Advances in Neural Information Processing Systems*, 36:20482–20494, 2023.

[195] Jiajun Deng, Tianyu He, Li Jiang, Tianyu Wang, Feras Dayoub, and Ian Reid. 3d-llava: Towards generalist 3d lmms with omni superpoint transformer, 2025.

[196] Yunze Man, Liang-Yan Gui, and Yu-Xiong Wang. Situational awareness matters in 3d vision language reasoning. In *CVPR*, 2024.

[197] Hongyan Zhi, Peihao Chen, Junyan Li, Shuailei Ma, Xinyu Sun, Tianhang Xiang, Yinjie Lei, Mingkui Tan, and Chuang Gan. Lscenellm: Enhancing large 3d scene understanding using adaptive visual preferences, 2025.

[198] Abdarahmane Traore, Éric Hervet, and Andy Couturier. Smolrgpt: Efficient spatial reasoning for warehouse environments with 600m parameters. *arXiv preprint arXiv:2509.15490*, 2025.

[199] Yang Liu, Ming Ma, Xiaomin Yu, Pengxiang Ding, Han Zhao, Mingyang Sun, Siteng Huang, and Donglin Wang. Ssr: Enhancing depth perception in vision-language models via rationale-guided spatial reasoning. *arXiv preprint arXiv:2505.12448*, 2025.

[200] 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. *arXiv preprint arXiv:2509.17664*, 2025.

[201] Zhiwen Fan, Jian Zhang, Renjie Li, Junge Zhang, Runjin Chen, Hezhen Hu, Kevin Wang, Huaizhi Qu, Dilin Wang, Zhicheng Yan, et al. Vlm-3r: Vision-language models augmented with instruction-aligned 3d reconstruction. *arXiv preprint arXiv:2505.20279*, 2025.

[202] Diankun Wu, Fangfu Liu, Yi-Hsin Hung, and Yueqi Duan. Spatial-mllm: Boosting mllm capabilities in visual-based spatial intelligence. *arXiv preprint arXiv:2505.23747*, 2025.

[203] Qianqian Wang*, Yifei Zhang*, Aleksander Holynski, Alexei A. Efros, and Angjoo Kanazawa. Continuous 3d perception model with persistent state. In *CVPR*, 2025.

[204] Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 5294–5306, 2025.

[205] Yining Hong, Chunru Lin, Yilun Du, Zhenfang Chen, Joshua B Tenenbaum, and Chuang Gan. 3d concept learning and reasoning from multi-view images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9202–9212, 2023.

[206] Anh Thai, Songyou Peng, Kyle Genova, Leonidas Guibas, and Thomas Funkhouser. Splattalk: 3d vqa with gaussian splatting. *arXiv preprint arXiv:2503.06271*, 2025.

[207] Zhangyang Qi, Zhixiong Zhang, Ye Fang, Jiaqi Wang, and Hengshuang Zhao. Gpt4scene: Understand 3d scenes from videos with vision-language models. *arXiv preprint arXiv:2501.01428*, 2025.

[208] Rao Fu, Jingyu Liu, Xilun Chen, Yixin Nie, and Wenhan Xiong. Scene-llm: Extending language model for 3d visual understanding and reasoning. *arXiv preprint arXiv:2403.11401*, 2024.

[209] Haifeng Huang, Yilun Chen, Zehan Wang, Rongjie Huang, Runsen Xu, Tai Wang, Luping Liu, Xize Cheng, Yang Zhao, Jiangmiao Pang, et al. Chat-scene: Bridging 3d scene and large language models with object identifiers. *arXiv preprint arXiv:2312.08168*, 2023.

[210] Weitai Kang, Haifeng Huang, Yuzhang Shang, Mubarak Shah, and Yan Yan. Robin3d: Improving 3d large language model via robust instruction tuning, 2025.

[211] Hanxun Yu, Wentong Li, Song Wang, Junbo Chen, and Jianke Zhu. Inst3d-lmm: Instance-aware 3d scene understanding with multi-modal instruction tuning, 2025.

[212] Jingzhou Luo, Yang Liu, Weixing Chen, Zhen Li, Yaowei Wang, Guanbin Li, and Liang Lin. Dspnet: Dual-vision scene perception for robust 3d question answering, 2025.

[213] Jiangyong Huang, Silong Yong, Xiaojian Ma, Xiongkun Linghu, Puhao Li, Yan Wang, Qing Li, Song-Chun Zhu, Baoxiong Jia, and Siyuan Huang. An embodied generalist agent in 3d world. *arXiv preprint arXiv:2311.12871*, 2023.

[214] Chenming Zhu, Tai Wang, Wenwei Zhang, Kai Chen, and Xihui Liu. Scanreason: Empowering 3d visual grounding with reasoning capabilities. In *European Conference on Computer Vision*, pages 151–168. Springer, 2024.

[215] Erik Daxberger, Nina Wenzel, David Griffiths, Haiming Gang, Justin Lazarow, Gefen Kohavi, Kai Kang, Marcin Eichner, Yinfei Yang, Afshin Dehghan, et al. Mm-spatial: Exploring 3d spatial understanding in multimodal llms. *arXiv preprint arXiv:2503.13111*, 2025.

[216] 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 *European conference on computer vision*, pages 740–755. Springer, 2014.

[217] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proc. Computer Vision and Pattern Recognition (CVPR), IEEE*, 2017.

[218] Weiyun Wang, Yiming Ren, Haowen Luo, Tiantong Li, Chenxiang Yan, Zhe Chen, Wenhai Wang, Qingyun Li, Lewei Lu, Xizhou Zhu, et al. The all-seeing project v2: Towards general relation comprehension of the open world. In *European Conference on Computer Vision*, pages 471–490. Springer, 2024.

[219] Haojun Jiang, Yuanze Lin, Dongchen Han, Shiji Song, and Gao Huang. Pseudo-q: Generating pseudo language queries for visual grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15513–15523, 2022.

[220] remyxai. Vqasynth, 2024. URL https://github.com/remyxai/VQASynth/tree/main. GitHub repository.

[221] Mingjie Xu, Mengyang Wu, Yuzhi Zhao, Jason Chun Lok Li, and Weifeng Ou. Llava-spacesgg: Visual instruct tuning for open-vocabulary scene graph generation with enhanced spatial relations, 2024.

[222] Shun-Cheng Wu, Johanna Wald, Keisuke Tateno, Nassir Navab, and Federico Tombari. Scenegraphfusion: Incremental 3d scene graph prediction from rgb-d sequences. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7515–7525, 2021.

[223] Xiongkun Linghu, Jiangyong Huang, Xuesong Niu, Xiaojian Shawn Ma, Baoxiong Jia, and Siyuan Huang. Multi-modal situated reasoning in 3d scenes. *Advances in Neural Information Processing Systems*, 37:140903–140936, 2024.

[224] Runsen Xu, Weiyao Wang, Hao Tang, Xingyu Chen, Xiaodong Wang, Fu-Jen Chu, Dahua Lin, Matt Feiszli, and Kevin J. Liang. Multi-spatialmllm: Multi-frame spatial understanding with multi-modal large language models. *arXiv preprint arXiv:2505.17015*, 2025.

[225] 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.

[226] 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.

[227] Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. Objaverse: A universe of annotated 3d objects. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13142–13153, 2023.

[228] Weichen Zhang, Zile Zhou, Zhiheng Zheng, Chen Gao, Jinqiang Cui, Yong Li, Xinlei Chen, and Xiao-Ping Zhang. Open3dvqa: A benchmark for comprehensive spatial reasoning with multimodal large language model in open space, 2025.

[229] Chen Gao, Baining Zhao, Weichen Zhang, Jinzhu Mao, Jun Zhang, Zhiheng Zheng, Fanhang Man, Jianjie Fang, Zile Zhou, Jinqiang Cui, et al. Embodiedcity: A benchmark platform for embodied agent in real-world city environment. *arXiv preprint arXiv:2410.09604*, 2024.

[230] Ji Hyeok Jung, Eun Tae Kim, Seoyeon Kim, Joo Ho Lee, Bumsoo Kim, and Buru Chang. Is 'right' right? enhancing object orientation understanding in multimodal large language models through egocentric instruction tuning, 2025.

[231] Shehreen Azad, Yash Jain, Rishit Garg, Yogesh S Rawat, and Vibhav Vineet. Understanding depth and height perception in large visual-language models, 2025.

[232] Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. Seed-bench: Benchmarking multimodal llms with generative comprehension, 2023.

[233] Wufei Ma, Haoyu Chen, Guofeng Zhang, Yu-Cheng Chou, Celso M de Melo, and Alan Yuille. 3dsrbench: A comprehensive 3d spatial reasoning benchmark. *arXiv preprint arXiv:2412.07825*, 2024.

[234] xAI Team. Grok-1.5v vision preview. https://x.ai/news/grok-1.5v, 2024. Accessed: 2024-07-07.

[235] Mengdi Jia, Zekun Qi, Shaochen Zhang, Wenyao Zhang, Xinqiang Yu, Jiawei He, He Wang, and Li Yi. Omnispatial: Towards comprehensive spatial reasoning benchmark for vision language models, 2025.

[236] OpenAI. Gpt-5 system card. Technical report, OpenAI, August 2025. Accessed: 2025-10-12.

[237] 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.

[238] Qwen Team. Qwen2.5-vl, January 2025. URL https://qwenlm.github.io/blog/qwen2.5-vl/.

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

[240] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llavanext: Improved reasoning, ocr, and world knowledge, 2024.

[241] 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.

[242] Jierui Peng, Yanyan Zhang, Yicheng Duan, Tuo Liang, Vipin Chaudhary, and Yu Yin. Nebula: Do we evaluate vision-language-action agents correctly?, 2025.

[243] Peiyan Li, Yixiang Chen, Hongtao Wu, Xiao Ma, Xiangnan Wu, Yan Huang, Liang Wang, Tao Kong, and Tieniu Tan. Bridgevla: Input-output alignment for efficient 3d manipulation learning with vision-language models. *arXiv preprint arXiv:2506.07961*, 2025.

[244] Tao Lin, Gen Li, Yilei Zhong, Yanwen Zou, Yuxin Du, Jiting Liu, Encheng Gu, and Bo Zhao. Evo-0: Vision-language-action model with implicit spatial understanding. *arXiv preprint arXiv:2507.00416*, 2025.

[245] Haoyu Zhen, Xiaowen Qiu, Peihao Chen, Jincheng Yang, Xin Yan, Yilun Du, Yining Hong, and Chuang Gan. 3d-vla: A 3d vision-language-action generative world model. *arXiv preprint arXiv:2403.09631*, 2024.

[246] Chengmeng Li, Junjie Wen, Yan Peng, Yaxin Peng, Feifei Feng, and Yichen Zhu. Pointvla: Injecting the 3d world into vision-language-action models. *arXiv preprint arXiv:2503.07511*, 2025.

[247] Delin Qu, Haoming Song, Qizhi Chen, Yuanqi Yao, Xinyi Ye, Yan Ding, Zhigang Wang, JiaYuan Gu, Bin Zhao, Dong Wang, and Xuelong Li. Spatialvla: Exploring spatial representations for visual-language-action model, 2025.

[248] Daniel Ekpo, Mara Levy, Saksham Suri, Chuong Huynh, and Abhinav Shrivastava. Verigraph: Scene graphs for execution verifiable robot planning, 2024.

[249] Nurhan Bulus Guran, Hanchi Ren, Jingjing Deng, and Xianghua Xie. Task-oriented robotic manipulation with vision language models, 2025.

[250] Guangyao Zhai, Xiaoni Cai, Dianye Huang, Yan Di, Fabian Manhardt, Federico Tombari, Nassir Navab, and Benjamin Busam. Sg-bot: Object rearrangement via coarse-to-fine robotic imagination on scene graphs, 2024.

[251] Helong Huang, Min Cen, Kai Tan, Xingyue Quan, Guowei Huang, and Hong Zhang. Graphcot-vla: A 3d spatial-aware reasoning vision-language-action model for robotic manipulation with ambiguous instructions, 2025.

[252] Yi Zhang, Qiang Zhang, Xiaozhu Ju, Zhaoyang Liu, Jilei Mao, Jingkai Sun, Jintao Wu, Shixiong Gao, Shihan Cai, Zhiyuan Qin, et al. Embodiedvsr: Dynamic scene graph-guided chain-of-thought reasoning for visual spatial tasks. *arXiv preprint arXiv:2503.11089*, 2025.

[253] Zhenhua Xu, Yujia Zhang, Enze Xie, Zhen Zhao, Yong Guo, Kwan-Yee K Wong, Zhenguo Li, and Hengshuang Zhao. Drivegpt4: Interpretable end-to-end autonomous driving via large language model. *IEEE Robotics and Automation Letters*, 2024.

[254] Zhenhua Xu, Yan Bai, Yujia Zhang, Zhuoling Li, Fei Xia, Kwan-Yee K Wong, Jianqiang Wang, and Hengshuang Zhao. Drivegpt4-v2: Harnessing large language model capabilities for enhanced closed-loop autonomous driving. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 17261–17270, 2025.

[255] Kexin Tian, Jingrui Mao, Yunlong Zhang, Jiwan Jiang, Yang Zhou, and Zhengzhong Tu. Nuscenes-spatialqa: A spatial understanding and reasoning benchmark for vision-language models in autonomous driving. *arXiv preprint arXiv:2504.03164*, 2025.

[256] Zhiyuan Zhang, Xiaofan Li, Zhihao Xu, Wenjie Peng, Zijian Zhou, Miaojing Shi, and Shuangping Huang. Mpdrive: Improving spatial understanding with marker-based prompt learning for autonomous driving. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 12089–12099, 2025.

[257] Shuang Zeng, Xinyuan Chang, Mengwei Xie, Xinran Liu, Yifan Bai, Zheng Pan, Mu Xu, and Xing Wei. Futuresightdrive: Thinking visually with spatio-temporal cot for autonomous driving, 2025.

[258] Ming Nie, Renyuan Peng, Chunwei Wang, Xinyue Cai, Jianhua Han, Hang Xu, and Li Zhang. Reason2drive: Towards interpretable and chain-based reasoning for autonomous driving. In *European Conference on Computer Vision*, pages 292–308. Springer, 2024.

[259] Oscar Michel, Anand Bhattad, Eli VanderBilt, Ranjay Krishna, Aniruddha Kembhavi, and Tanmay Gupta. Object 3dit: Language-guided 3d-aware image editing, 2023.

[260] Tsung-Han Wu, Long Lian, Joseph E. Gonzalez, Boyi Li, and Trevor Darrell. Self-correcting llm-controlled diffusion models, 2023.

[261] Jaemin Cho, Abhay Zala, and Mohit Bansal. Dall-eval: Probing the reasoning skills and social biases of text-to-image generation models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3043–3054, 2023.

[262] Yi Huang, Jiancheng Huang, Yifan Liu, Mingfu Yan, Jiaxi Lv, Jianzhuang Liu, Wei Xiong, He Zhang, Liangliang Cao, and Shifeng Chen. Diffusion model-based image editing: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

[263] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18392–18402, 2023.

[264] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022.

[265] Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. Plug-and-play diffusion features for text-driven image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1921–1930, 2023.

[266] Xiaofeng Wang, Zheng Zhu, Guan Huang, Xinze Chen, Jiagang Zhu, and Jiwen Lu. Drivedreamer: Towards real-world-driven world models for autonomous driving, 2023.

[267] Xindi Yang, Baolu Li, Yiming Zhang, Zhenfei Yin, Lei Bai, Liqian Ma, Zhiyong Wang, Jianfei Cai, Tien-Tsin Wong, Huchuan Lu, and Xu Jia. Vlipp: Towards physically plausible video generation with vision and language informed physical prior, 2025.

[268] Yanzhao Shi, Xiaodan Zhang, Junzhong Ji, Haoning Jiang, Chengxin Zheng, Yinong Wang, and Liangqiong Qu. Hsenet: Hybrid spatial encoding network for 3d medical vision-language understanding, 2025.

[269] Changsun Lee, Sangjoon Park, Cheong-Il Shin, Woo Hee Choi, Hyun Jeong Park, Jeong Eun Lee, and Jong Chul Ye. Read like a radiologist: Efficient vision-language model for 3d medical imaging interpretation, 2024.

[270] Yu Xin, Gorkem Can Ates, Kuang Gong, and Wei Shao. Med3dvlm: An efficient vision-language model for 3d medical image analysis, 2025.

[271] Lin Duan, Yanming Xiu, and Maria Gorlatova. Advancing the understanding and evaluation of ar-generated scenes: When vision-language models shine and stumble, 2025.

[272] Xiangzhi Eric Wang, Zackary P. T. Sin, Ye Jia, Daniel Archer, Wynonna H. Y. Fong, Qing Li, and Chen Li. Can you move these over there? an llm-based vr mover for supporting object manipulation, 2025.

[273] Jiahuan Pei, Irene Viola, Haochen Huang, Junxiao Wang, Moonisa Ahsan, Fanghua Ye, Jiang Yiming, Yao Sai, Di Wang, Zhumin Chen, et al. Autonomous workflow for multimodal fine-grained training assistants towards mixed reality. *arXiv preprint arXiv:2405.13034*, 2024.

[274] Tianxiao Zhang, Wenju Xu, Bo Luo, and Guanghui Wang. Depth-wise convolutions in vision transformers for efficient training on small datasets. *Neurocomputing*, 617:128998, 2025.

[275] Elia Peruzzo, Enver Sangineto, Yahui Liu, Marco De Nadai, Wei Bi, Bruno Lepri, and Nicu Sebe. Spatial entropy as an inductive bias for vision transformers, 2023.

[276] Itamar Zimerman and Lior Wolf. Multi-dimensional hyena for spatial inductive bias, 2023.

[277] Yuxuan Zhou, Wangmeng Xiang, Chao Li, Biao Wang, Xihan Wei, Lei Zhang, Margret Keuper, and Xiansheng Hua. Sp-vit: Learning 2d spatial priors for vision transformers, 2022.

[278] Mengfei Du, Binhao Wu, Zejun Li, Xuanjing Huang, and Zhongyu Wei. Embspatial-bench: Benchmarking spatial understanding for embodied tasks with large vision-language models. *arXiv preprint arXiv:2406.05756*, 2024.

[279] 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 *European Conference on Computer Vision*, pages 148–166. Springer, 2024.

[280] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, and Rongrong Ji. Mme: A comprehensive evaluation benchmark for multimodal large language models, 2024.

[281] Mohsen Gholami, Ahmad Rezaei, Zhou Weimin, Sitong Mao, Shunbo Zhou, Yong Zhang, and Mohammad Akbari. Spatial reasoning with vision-language models in ego-centric multi-view scenes, 2025.

[282] Ziyang Gong, Wenhao Li, Oliver Ma, Songyuan Li, Jiayi Ji, Xue Yang, Gen Luo, Junchi Yan, and Rongrong Ji. Space-10: A comprehensive benchmark for multimodal large language models in compositional spatial intelligence. *arXiv preprint arXiv:2506.07966*, 2025.

[283] Xianda Guo, Ruijun Zhang, Yiqun Duan, Yuhang He, Dujun Nie, Wenke Huang, Chenming Zhang, Shuai Liu, Hao Zhao, and Long Chen. Surds: Benchmarking spatial understanding and reasoning in driving scenarios with vision language models, 2025.

[284] Nils Hoehing, Ellen Rushe, and Anthony Ventresque. What's left can't be right–the remaining positional incompetence of contrastive vision-language models. *arXiv preprint arXiv:2311.11477*, 2023.

[285] Ji Hyeok Jung, Eun Tae Kim, Seo Yeon Kim, Joo Ho Lee, Bumsoo Kim, and Buru Chang. Is' right'right? enhancing object orientation understanding in multimodal language models through egocentric instruction tuning. *arXiv preprint arXiv:2411.16761*, 2024.

[286] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. ReferItGame: Referring to objects in photographs of natural scenes. In Alessandro Moschitti, Bo Pang, and Walter Daelemans, editors, *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 787–798, Doha, Qatar, October 2014. Association for Computational Linguistics. doi: 10.3115/v1/D14-1086.

[287] 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 *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4015–4026, 2023.

[288] Mengcheng Lan, Chaofeng Chen, Yiping Ke, Xinjiang Wang, Litong Feng, and Wayne Zhang. Proxyclip: Proxy attention improves clip for open-vocabulary segmentation. In *European Conference on Computer Vision*, pages 70–88. Springer, 2024.

[289] Jianing Li, Xi Nan, Ming Lu, Li Du, and Shanghang Zhang. Proximity qa: Unleashing the power of multi-modal large language models for spatial proximity analysis. *arXiv preprint arXiv:2401.17862*, 2024.

[290] Yun Li, Yiming Zhang, Tao Lin, Xiangrui Liu, Wenxiao Cai, Zheng Liu, and Bo Zhao. Sti-bench: Are mllms ready for precise spatial-temporal world understanding?, 2025.

[291] Dingming Li, Hongxing Li, Zixuan Wang, Yuchen Yan, Hang Zhang, Siqi Chen, Guiyang Hou, Shengpei Jiang, Wenqi Zhang, Yongliang Shen, Weiming Lu, and Yueting Zhuang. Viewspatial-bench: Evaluating multi-perspective spatial localization in vision-language models, 2025.

[292] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pages 216–233. Springer, 2024.

[293] Xiaojian Ma, Silong Yong, Zilong Zheng, Qing Li, Yitao Liang, Song-Chun Zhu, and Siyuan Huang. Sqa3d: Situated question answering in 3d scenes. *arXiv preprint arXiv:2210.07474*, 2022.

[294] Margherita Malanchini, Kaili Rimfeld, Nicholas G Shakeshaft, Andrew McMillan, Kerry L Schofield, Maja Rodic, Valerio Rossi, Yulia Kovas, Philip S Dale, Elliot M Tucker-Drob, et al. Evidence for a unitary structure of spatial cognition beyond general intelligence. *npj Science of Learning*, 5(1):9, 2020.

[295] Damiano Marsili, Rohun Agrawal, Yisong Yue, and Georgia Gkioxari. Visual agentic ai for spatial reasoning with a dynamic api. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pages 19446–19455, June 2025.

[296] Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. Kosmos-2: Grounding multimodal large language models to the world. *arXiv preprint arXiv:2306.14824*, 2023.

[297] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pages 2641–2649, 2015.

[298] Congpei Qiu, Yanhao Wu, Wei Ke, Xiuxiu Bai, and Tong Zhang. Refining clip's spatial awareness: A visual-centric perspective. *arXiv preprint arXiv:2504.02328*, 2025.

[299] Santhosh Kumar Ramakrishnan, Erik Wijmans, Philipp Kraehenbuehl, and Vladlen Koltun. Does spatial cognition emerge in frontier models? In *International Conference on Learning Representations*, 2025.

[300] 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, 2025.

[301] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2022.

[302] Lin Sun, Jiale Cao, Jin Xie, Xiaoheng Jiang, and Yanwei Pang. Cliper: Hierarchically improving spatial representation of clip for open-vocabulary semantic segmentation, 2024.

[303] Emilia Szymańska, Mihai Dusmanu, Jan-Willem Buurlage, Mahdi Rad, and Marc Pollefeys. Space3d-bench: Spatial 3d question answering benchmark. In *European Conference on Computer Vision*, pages 68–85. Springer, 2025.

[304] Kevis-Kokitsi Maninis, Kaifeng Chen, Soham Ghosh, Arjun Karpur, Koert Chen, Ye Xia, Bingyi Cao, Daniel Salz, Guangxing Han, Jan Dlabal, Dan Gnanapragasam, Mojtaba Seyedhosseini, Howard Zhou, and André Araujo. TIPS: Text-Image Pretraining with Spatial Awareness. In *ICLR*, 2025.

[305] Xingrui Wang, Wufei Ma, Zhuowan Li, Adam Kortylewski, and Alan Yuille. 3d-aware visual question answering about parts, poses and occlusions, 2023.

[306] Haoxiang Wang, Pavan Kumar Anasosalu Vasu, Fartash Faghri, Raviteja Vemulapalli, Mehrdad Farajtabar, Sachin Mehta, Mohammad Rastegari, Oncel Tuzel, and Hadi Pouransari. Sam-clip: Merging vision foundation models towards semantic and spatial understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3635–3647, 2024.

[307] Feng Wang, Jieru Mei, and Alan Yuille. Sclip: Rethinking self-attention for dense vision-language inference. In *European Conference on Computer Vision*, pages 315–332. Springer, 2024.

[308] Wenqi Wang, Reuben Tan, Pengyue Zhu, Jianwei Yang, Zhengyuan Yang, Lijuan Wang, Andrey Kolobov, Jianfeng Gao, and Boqing Gong. Site: towards spatial intelligence thorough evaluation, 2025.

[309] Haoning Wu, Xiao Huang, Yaohui Chen, Ya Zhang, Yanfeng Wang, and Weidi Xie. Spatialscore: Towards unified evaluation for multimodal spatial understanding, 2025.

[310] Linhui Xiao, Xiaoshan Yang, Xiangyuan Lan, Yaowei Wang, and Changsheng Xu. Towards visual grounding: A survey. *arXiv preprint arXiv:2412.20206*, 2024.

[311] Sihan Yang, Runsen Xu, Yiman Xie, Sizhe Yang, Mo Li, Jingli Lin, Chenming Zhu, Xiaochen Chen, Haodong Duan, Xiangyu Yue, Dahua Lin, Tai Wang, and Jiangmiao Pang. Mmsi-bench: A benchmark for multi-image spatial intelligence, 2025.

[312] Chun-Hsiao Yeh, Chenyu Wang, Shengbang Tong, Ta-Ying Cheng, Ruoyu Wang, Tianzhe Chu, Yuexiang Zhai, Yubei Chen, Shenghua Gao, and Yi Ma. Seeing from another perspective: Evaluating multi-view understanding in mllms. *arXiv preprint arXiv:2504.15280*, 2025.

[313] Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*, 2023.

[314] Yuyou Zhang, Radu Corcodel, Chiori Hori, Anoop Cherian, and Ding Zhao. Spinbench: Perspective and rotation as a lens on spatial reasoning in vlms. *arXiv preprint arXiv:2509.25390*, 2025.

[315] Yiqi Zhu, Ziyue Wang, Can Zhang, Peng Li, and Yang Liu. Cospace: Benchmarking continuous space perception ability for vision-language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 29569–29579, 2025.

A Appendix Organization

In the appendix, we provide the supplementary materials for this survey, including

- the dataset and benchmark collection in Sec. A.1, as mentioned in the main paper.
- implementation details of evaluation in Sec. A.2.

A.1 Dataset and Benchmarks Collection

Reviewing the impressive progress in spatial capabilities of VLMs, we collected datasets used for training and benchmarks used for assessing spatial reasoning in VLMs over the past two years. As shown in Tabs. 2 and 3, we list these datasets along with their publication venues and categorize them into corresponding cognitive levels based on the predominant types of QA pairs they contain. The "Fundamental Task" column specifies the concrete tasks within each dataset, and the "Size" column represents the number of QA pairs. The "Image Source" column indicates where the original visual data originate from, while the "Modality" column specifies the types of visual modalities included in each corpus.

A.2 Implementation Details

We provide implementation details in § 6. Sec.A.2.1 describes the benchmarks used in evaluation, and Sec.A.2.2 presents the configuration details for model inference.

A.2.1 Evaluation Dataset Descriptions

In this section, we present brief descriptions of the datasets used in the Tab. 1. To rigorously assess spatial capabilities across distinct cognitive levels—perception, understanding, and extrapolation—we curate tailored subsets from the original datasets, denoted with *:

- **EgoOrientBench***: a benchmark that evaluates MLLMs' orientation understanding across three subsets. In our benchmarking, we use the *choose* and *verify* subsets, which contain a single ground truth, thus avoiding the ambiguity that may arise in the *freeform* subset.
- **GeoMeter***: a benchmark designed to evaluate perception of object dimensions and scene depth across both real-world and synthetic settings. In our benchmarking, we use only the *real-world* subset, avoiding the noisy design in synthetic setting.
- **CV-Bench***: a vision-centric benchmark focusing on 2D and 3D visual understanding, consisting of four main question types: spatial relationship, object counting, depth order, and relative distance. In our benchmarking, we use the *spatial relationship* and *relative distance* subsets to evaluate the spatial understanding of VLMs.
- **What's Up**: a benchmark consisting of sets of photographs that vary only in the spatial relations between objects while keeping their identities fixed.
- **SEED-Bench***: a benchmark spanning 12 evaluation dimensions, including general comprehension of both image and video modalities. In our benchmarking, we use the *spatial relation* and *instance localization* subsets to gauge spatial understanding between objects under the original **Spatial Understanding** category.
- **SRBench***: a benchmark aimed at evaluating spatial reasoning through its core components, including spatial relations, orientation and navigation, mental rotation, and spatial visualization. In our benchmarking, we exclusively use data from the following subsets: *Mental Rotation Tests (MRT)*, *Paper Fold*, and *Navigation*, to focus on extrapolation measurement.
- **MINDCUBE**: a benchmark that evaluates VLMs from the perspective of spatial mental modeling and internal representation. We use its multi-view image–question pairs to assess high-level extrapolative reasoning.
- **RealWorldQA**: a benchmark dataset released by xAI, designed to evaluate the basic real-world spatial understanding capabilities of multimodal models. We use this dataset for overall assessment, as it covers all three levels of spatial cognition defined in this paper.
- **OmniSpatial**: a benchmark covering four major categories, including dynamic reasoning, complex spatial logic, spatial interaction, and perspective-taking, with 50 fine-grained subcategories. Similar to RealWorldQA, we use this dataset for overall evaluation due to its comprehensive coverage.

A.2.2 Model Inference Configuration

As shown in Tab.4, we list all the models evaluated in Tab.1, along with their corresponding details. The category of each model remains consistent with those presented in Tab. 1. The model names and specific versions from the source such as Huggingface, Github are also provided in Tab.4. In addition, we indicate the model backbone for each work, excluding general-purpose VLMs which it is more easy to see the comparison between backbone model and its related improved model, and specify whether the models support multiview input. During evaluation, the *max_new_tokens* parameter is set to *1024*, and *do_sample* is set to *False* for all models to ensure consistent and deterministic results.

Table 2 A collection of 21 spatial training datasets published between January 2023 and October 2025. **P.**, **U.**, and **E.** denote the cognitive levels of *perception*, *understanding*, and *extrapolation*, respectively. See Fig. 3 for dataset volume trend.

Dataset	Venue	P.	U.	E.	Fundamental Task	Size	Image Source	Modality
Proximity-110K [289]	ArXiv2024	✓			depth estimation	989,877	Visual Genome, COCO	RGB
AS-V2 [218]	ECCV2024		✓		Spatial Relations VQA	127,000	COCO	RGB
SpaceThinker [220]	online		✓		Spatial Relations VQA	12,000	VQASynth	RGB
OpenSpatial [67]	NeurIPS2024		✓		Spatial Relations VQA	8,700,000	OpenImages	RGB-D
SUN-Spot v2.0 [116]	ArXiv2025		✓		Spatial Relations VQA	101,053	SUN RGB-D	RGB-D
SQA3D [293]	ICLR2023			✓	Spatial Situated Reasoning	33,400	ScanNet	RGB, Point Cloud
MSQA [223]	NeurIPS2024			✓	Spatial Situated Reasoning	251,000	ScanNet , 3RScan , ARKitScenes	Point Cloud
Spartun3D [123]	ICLR2025			✓	Spatial Situated Reasoning	123,000	3RScan	Point Cloud
MulSeT [94]	ArXiv2025			✓	Spatial Situated Reasoning, Spatial Simulation and Inferring	38,200	AI2THOR	RGB
Super-CLEVR-3D(Pose&occlusion) [305]	NeurIPS2023	✓	✓		Orientation Estimation, Spatial relation VQA	543,383	Super-CLEVR	RGB
SpatialQA [15]	ICRA2025	✓	✓		Depth estimation, 3D object detection, Spatial relation VQA	852,869	Bunny 695k, Open X-Embodiment	RGB-D
SURDS [283]	ArXiv2025	✓	✓		Depth estimation, Orientation estiamtion, Spatial Relations VQA	50,330	nuScenes	RGB
Open3DVQA[228]	ArXiv2025	✓	✓		3D Object Detection, Orientation estimation, Spatial Relations VQA,	9,048	synthetic data	RGB-D
RefSpatial [17]	NeurIPS2025	✓	✓		3D Object Detection, Depth estimation, Spatial relation QA	22,000,000	OpenImages, CA-1M, synthetic	RGB-D
SSR-CoT [199]	NeurIPS2025	✓	✓		3D Object Detection, Depth estimation, Spatial relation QA	1,200,000	LLaVA-CoT, Visual-CoT, VoCoT, SpatialQA	RGB-D

(Continued from TABLE 2)

Dataset	Venue	P.	U.	E.	Fundamental Task	Size	Image Source	Modality
SR-91k [155]	ArXiv2025	✓		✓	Depth estimation, Object Detection, Spatial Situated Reasoning	91,000	ScanNet	RGB
SpaceSGG[221]	WACV2025		✓	✓	Spatial Relation VQA, Spatial Situated Reasoning	40,000	COCO	RGB
SAT [300]	ArXiv2025	✓	✓	✓	Depth estimation,Spatial relations VQA, Spatial situated reasoning, Spatial Simulation and Inferring	175,000	PixMo-Cap, DOCCI, Pixmo-Cap	RGB
SPAR-7M [18]	ArXiv2025	✓	✓	✓	Depth estimation, Spatial relations VQA, Spatial simulation and inferring, Spatial situated reasoning	7,000,000	ProcTHOR-10K	RGB
RoboSpatial [19]	CVPR2025	✓	✓	✓	3D Object Detection, Spatial Relations VQA, Spatial Situated Reasoning, Spatial Simulation and Inferring	3,000,000	Matterport3D, ScanNet, 3RScan, HOPE, GraspNet-1B	RGB, Point Cloud
MSMU [200]	NeurIPS2025	✓	✓	✓	3D Object Detection, Spatial Relations VQA, Spatial Simulation and Inferring	700,000	ScanNet, ScanNet++,	RGB

Table 3 A collection of 49 spatial benchmarks published between January 2023 and October 2025. **P.**, **U.**, and **E.** denote the cognitive levels of *perception*, *understanding*, and *extrapolation*, respectively. See Fig. 4 for dataset volume trend.

Dataset	Venue	P.	U.	E.	Fundamental Task	Size	Image Source	Modality
Ori-Bench[226]	ArXiv2024	✓			orientation estimation	400	COCO, Generated from DALL-E	RGB
EgoOrientBench[285]	CVPR2025	✓			orientation estimation	33,460	ImageNet, D3, DomainNet, PACS, OmniObject3D	RGB
GeoMeter [231]	CVPRW2025	✓			depth estimation	11,200	synthetic data	RGB
CRPE-relation[218]	ECCV2024		✓		Spatial Relations VQA	7,576	COCO	RGB
MMBench(physical, spatial relation)[292]	ArXiv2024		✓		Spatial Relations VQA	251	Online	RGB
SEED-Bench (Spatial Rel.&Instance Loc.)[232]	CVPR2024		✓		Spatial Relations VQA	1,634	CC3M	RGB
MM-Vet(Spatial awareness (Spat))[313]	ICML2024		✓		Spatial Relations VQA	75	Online	RGB
TopViewRS[83]	EMNLP2024		✓		spatial relation VQA	11,384	Matterport3D	RGB, 3D Mesh
MME(position split)[280]	ArXiv2024		✓		spatial relation VQA	60	COCO	RGB
EmbSpatial-Bench[278]	ACL2024		✓		spatial relation VQA	3,640	MP3D, AI2-THOR, ScanNet	RGB
What's Up[51]	EMNLP2023		✓		spatial relation VQA	4,138	COCO,GQA	RGB
VSP[75]	ArXiv2024			✓	Spatial Situated Reasoning	4,600	OpenAI Gym, BIRD	RGB
MapBench[122]	ArXiv2025			✓	Spatial Situated Reasoning	1,649	online	RGB
MINDCUBE[26]	ArXiv2025			✓	Spatial Situated Reasoning	21,154	ArkitScenes,DL3DV-10K,WildRGB-D	RGB
MMSI-Bench [311]	ArXiv2025			✓	Spatial Situated Reasoning	1,000	Matterport3D, ScanNet, ...	RGB
CoSpace[315]	CVPR2025			✓	Spatial Situated Reasoning, Spatial Simulation and Inferring	1,626	Baidu Map Panorama API, HM3D	RGB
SPACE-Visual[299]	ICLR2025			✓	Spatial Situated Reasoning, Spatial Simulation and Inferring	5,008	Synthetic Data, Oline	RGB
CV-Bench[13]	NeurIPS2024	✓	✓		Depth Estimation, Spatial Relation VQA	2,638	COCO,ADE20K, Omini3D	RGB
SpatialRGPT-Bench[67]	Neurips2024	✓	✓		3D Object Detection, Depth Estimation, Spatial Relations VQA,	1,410	nuScenes,Hypersim, SUNRGBD, KITTI ARKitScenes ...	RGB
Q-Spatial[12]	EMNLP2024	✓	✓		3D Object Detection, Spatial Relations VQA	271	ScanNet, images captured by iPhone	RGB

(Continued from TABLE 3)

Dataset	Venue	P.	U.	E.	Fundamental Task	Size	Image Source	Modality
SpatialBench[15]	ICRA2025	✓	✓		depth estimation, Spatial relations VQA, 3D object detection,	182	MME and manually annotated images	RGB-D
Omni3D-Bench[295]	CVPR2025	✓	✓		3D Object Detection, Spatial Relations VQA	500	Omni3D	RGB
STI-Bench[290]	ICCV2025	✓	✓		3D Object Detection, Spatial Relations, orientation estimation	2,060	Waymo, ScanNet, Omni6DPose	RGB
RefSpatial-Bench[17]	NeurIPS2025	✓	✓		3D Object Detection,Depth estimation,Spatial relation VQA	200	manually collect	RGB
SSRBENCH-Spatial[199]	NeurIPS2025	✓	✓		3D Object Detection,Spatial relation VQA	357	SSR-CoT	RGB-D
BLINK[279]	ECCV2024	✓		✓	depth estimation, Spatial simulation in inferring, Spatial situated reasoning	3,807	HPatches	RGB
All-Angles-Bench[312]	ArXiv2025	✓		✓	depth estimation, Spatial simulation in inferring, Spatial situated reasoning	2,132	Exo4D, EgoHumans	RGB
SpaceSGG-Val [221]	WACV2025		✓	✓	Spatial Simulation and Inferring, Spatial Situated Reasoning, Spatial relations VQA	271	COCO	RGB
Space3D-Bench(Relation, Navigation, Prediction)[303]	ECCV2024		✓	✓	Spatial Simulation and Inferring, Spatial Situated Reasoning, Spatial relations VQA	1,000	Replica	RGB, 3D Mesh
SpatialEval(VQA, VTQA)[47]	Neurips2024		✓	✓	Spatial situated reasoning, spatial relation VQA	9,270	synthetic, Densely Captioned Images (DCI)	RGB
3DSRBench[233]	ICCV2025		✓	✓	Spatial situated reasoning, spatial relation VQA	2,170	MS-COCO, HSSD	RGB
VSR[85]	ACL2023		✓	✓	Spatial situated reasoning, spatial relation VQA	10,972	MS-COCO	RGB
Spatial457[117]	CVPR2025		✓	✓	Spatial simulation in inferring, spatial relation VQA	23,752	synthetic data	RGB
COMFORT[82]	ICLR2025		✓	✓	Spatial situated reasoning, spatial relation VQA	58,320	synthetic data	RGB
SRBench[72]	ArXiv2025	✓	✓	✓	Orientation Estimation, Spatial Relations VQA, Spatial Simulation and Inferring	1,800	classic Mental Rotation Test, EgoOrientBench, Spatial-MM	RGB
VSI-Bench[118]	CVPR2025	✓	✓	✓	3D object detection, Spatial relations VQA, Spatial situated reasoning	5,130	ScanNet, ScanNet++, and ARKitScenes	RGB
RealWorldQA[234]	-	✓	✓	✓	depth estimation, Spatial relations VQA, Spatial simulation in inferring, Spatial situated reasoning	765	NA	RGB

(Continued from TABLE 3)

Dataset	Venue	P.	U.	E.	Fundamental Task	Size	Image Source	Modality
SPAR-Bench[18]	ArXiv2025	✓	✓	✓	depth estimation, Spatial relations VQA, Spatial simulation in inferring, Spatial situated reasoning	7,207	SPAR-7M	RGB
SPHERE[119]	ArXiv2025	✓	✓	✓	3D Object Detection, Spatial Relations VQA, Spatial Simulation and Inferring, Spatial Situated Reasoning	2,285	MS COCO-2017	RGB
ViewSpatial-Bench[291]	ArXiv2025	✓	✓	✓	orientation estimation, Spatial relations VQA, Spatial Situated Reasoning	5,700	MS-CoCo, ScanNet	RGB
Spatial-MM[121]	EMNLP2024	✓	✓	✓	Orientation Estimation, Spatial Relations VQA, Spatial Situated Reasoning	2,310	Onlin	RGB
SpatialScore[309]	ArXiv2025	✓	✓	✓	3D Object Detection, Spatial Relations VQA, Spatial Situated Reasoning, Spatial Simulation and Inferring	28,000	MMVP, MMIU, RealWorldQA , SpatialSense, SpatialBench, ...	RGB
OmniSpatial[235]	ArXiv2025	✓	✓	✓	3D Object Detection, Depth Estimation, orientation estimation, Spatial Relations VQA, Spatial Situated Reasoning, Spatial Simulation and Inferring	1,500	Web Images, Exam-Based Test Questions, Driving Test Questions, MME, HOI4D	RGB
SITE-Bench[308]	ArXiv2025	✓	✓	✓	3D Object Detection Spatial Relations VQA Spatial Situated Reasoning Spatial Simulation and Inferring	8,068	VSI-Bench, Blink, VSR, MMBench, ...	RGB
RoboSpatial-Home[19]	CVPR2025	✓	✓	✓	3D Object Detection Spatial Relations VQA Spatial Situated Reasoning Spatial Simulation and Inferring	350	manually collect	RGB-D
Ego3D-Bench[281]	ArXiv2025	✓	✓	✓	3D Object Detection,Spatial Relations,Spatial Simulation and Inferring, Spatial Situated Reasoning	8,600	manually collect	RGB
MSMU-Bench[200]	NeurIPS2025	✓	✓	✓	3D Object Detection, Spatial Relations VQA, Spatial Simulation and Inferring	1,000	ScanNet, ScanNet++	RGB-D
SPINBENCH[314]	ArXiv2025	✓	✓	✓	Depth Estimation, Spatial Relations VQA, Spatial Simulation and Inferring, Spatial Situated Reasoning	2,599	Synthetic data, Multi-View Car Dataset, Stereo Face Database	RGB
SPACE-10 [282]	ArXiv2025	✓	✓	✓	3D Object Detection, Spatial Relations VQA, Spatial Simulation and Inferring, Spatial Situated Reasoning	5,000	SCN, 3RS, ARK, SCN++	RGB, Point Cloud

Table 4 Comparison of general and spatially enhanced VLMs.

Type **Models / Methods**	**Model Version**	**Model Source**	**Model Backbone**	**Multi-View**
General Models				
GPT-4o	gpt-4o-2024-08-06	OpenAI	–	✓
GPT-5	gpt-5-2025-08-07	OpenAI	–	✓
Gemini 2.5 flash	gemini-2.5-flash	Google Cloud	–	✓
Gemini 2.5 pro	gemini-2.5-pro	Google Cloud	–	✓
Qwen2.5-VL-7B	Qwen/Qwen2.5-VL-7B-Instruct	Huggingface	–	✓
Qwen2.5-VL-72B	Qwen/Qwen2.5-VL-72B-Instruct	Huggingface	–	✓
LLaVA-v1.5-7B	llava-hf/llava-1.5-7b-hf	Huggingface	–	✓
LLaVA-NeXT-7B	llava-hf/llava-v1.6-mistral-7b-hf	Huggingface	–	✓
LLaVA-OneVision-7B	llava-hf/llava-onevision-qwen2-7b-ov-hf	Huggingface	–	✓
LLaVA-Next-72B	llava-hf/llava-next-72b-hf	Huggingface	–	✓
5.2 Model-Centric Enhancement				
ROSS	HaochenWang/ross-qwen2-7b	Huggingface	CLIP-ViT-L+Qwen2-7B	✓
ViLaSR	inclusionAI/ViLaSR	Huggingface	Qwen-2.5-VL-7B	✓
M2-Reasoning-7B	inclusionAI/M2-Reasoning	Huggingface	Qwen-2.5-VL-7B	✓
LLaVA-AURORA	LLaVA-AURORA	Github	LLaVA-v1.5-13B	✓
AdaptVis	llava_1.5_adapt_vis	Github	LLaVA-v1.5-7B	✓
Honeybee	Honeybee-C-7B-M256	Github	CLIP ViT-L+Vicuna v1.5-7B	✓
Cambrian-1	nyu-visionx/cambrian-8b	Huggingface	CLIP ViT-L+Vicuna-1.5-7B	✓
5.3 Explicit 2D Information Injecting				
VPT	rp-yu/Qwen2-VL-7b-VPT-CLIP	Huggingface	Qwen2-VL-7B	✓
VCoder	shi-labs/vcoder_llava-v1.5-7b	Huggingface	LLaVA-v1.5-7B	✓
5.4 3D Spatial Information Enhancement				
LLaVA-3D	ChaimZhu/LLaVA-3D-7B	Huggingface	LLaVA-v1.5-7B	✓
SpatialBot-3B	RussRobin/SpatialBot-3B	Huggingface	Phi2-3B	✗
VCoder (depth)	shi-labs/vcoder_ds_llava-v1.5-7b	Huggingface	Depth Encoder + LLaVA-v1.5-7B	✓
Data-Centric Spatial Enhancement				
SpaceOm	remyxai/SpaceOm	Huggingface	Qwen2.5VL-3B	✓
SpaceQwen2.5-VL-3B-Instruct	remyxai/SpaceQwen2.5-VL-3B-Instruct	Huggingface	Qwen2.5-VL-3B	✓
SpaceFlorence-2	remyxai/SpaceFlorence-2	Huggingface	Florence-2-base	✗
SpaceThinker-Qwen2.5VL-3B	remyxai/SpaceThinker-Qwen2.5VL-3B	Huggingface	Qwen2.5-VL-3B	✓
SpaceMantis	remyxai/SpaceMantis	Huggingface	Mantis-8B	✓
SpaceLLaVA-13B	remyxai/SpaceLLaVA	Huggingface	LLaVA-v1.5-13B	✓
SpaceLLaVA-1.5-7B	salma-remyx/spacellava-1.5-7b	Huggingface	LLaVA-v1.5-7B	✓