Spatial Reasoning in Multimodal Foundation Models: A Survey of Spatial Understanding in LLMs and VLMs

Azmine Toushik Wasi[1,2†], **Md. Masudur Rahman**[1], **Abdur Rahman**[1,2], **Mohsin Mahmud Topu**[1,2], **Mohammed Eunus Ali**[3], **Riashat Islam**[4], **Md Rizwan Parvez**[5]
[1]Computational Intelligence and Operations Laboratory
[2]Shahjalal University of Science and Technology
[3]Monash University [4]Microsoft AI Frontiers [5]Qatar Computing Research Institute
[†]**Correspondence:** azmine32@student.sust.edu

Abstract

Spatial reasoning, the ability to represent and reason about spatial relations, geometry, topology, and embodied environments, is increasingly vital for evaluating and deploying large language models (LLMs) and vision–language models (VLMs). While multimodal foundation models show strong performance on spatial benchmarks, navigation, and embodied planning, much of this stems from shallow pattern matching rather than grounded spatial understanding. This review offers a comprehensive, framework-driven analysis of spatial reasoning in LLMs and VLMs, covering tasks, representations, datasets, benchmarks, architectures, training, and failure modes. We also link spatial reasoning to embodied AI, emphasizing challenges in generalization, compositionality, and grounding. Our survey unifies diverse fields, NLP, vision, robotics, and cognitive science, to guide the development of truly spatially grounded foundation models.

1 Introduction

Spatial reasoning is fundamental to intelligent behavior, enabling agents to perceive, interpret, and act within structured environments, and is especially critical for vision-language models (VLMs) in embodied settings (Chen et al., 2024; Munje et al., 2025; Song et al., 2025a). For VLMs, spatial competence supports relational understanding, navigation, manipulation, and quantitative estimation, directly affecting reliability in real-world applications (Chen et al., 2024; Song et al., 2025a). Although early visual question answering suggested promising capabilities, later studies show that robust spatial understanding remains limited, particularly beyond simplified 2D scenes (Johnson et al., 2017; Hudson and Manning, 2019). This weakness is critical in robotics and embodied AI, where spatial inference failures can cause real-world errors (Feng et al., 2025; Ma et al., 2026).

LLMs and VLMs struggle with spatial reasoning because spatial concepts are continuous, multi-referential, and perspective-dependent, requiring grounded representations beyond symbolic patterns (Stogiannidis et al., 2025; Tasnim et al., 2026). Many models rely on 2D cues and dataset biases, leading to brittle performance under viewpoint shifts and occlusion (Pothiraj et al., 2025; Wang and Ling, 2025; Zha et al., 2025). Difficulties increase with multi-step relational composition and metric estimation, which remain challenging for autoregressive models (Chen et al., 2024; Jia et al., 2026a). Strong benchmark performance often fails to generalize to 3D reasoning or perspective-taking (Suhr et al., 2017), while embodied and navigation benchmarks expose limitations even in top models, suggesting reliance on pattern recognition rather than genuine world modeling (Danish et al., 2025; Song et al., 2025b).

In this survey, we review spatial reasoning in large language and vision–language models, focusing on the gap between benchmark performance and genuine spatial understanding. Drawing on cognitive science and geometry, we distinguish spatial reasoning from recognition and pattern matching(§2) and introduce a multi-axis framework (§3) spanning task demands, spatial relations, reasoning methods, representations, and temporal–memory needs across NLP, vision, and robotics. This framework exposes persistent gaps in grounding, metric consistency, and long-term inference, and links recurring failure modes to mitigations and residual limitations (Table 3; §4). The appendix covers evaluation protocols, architectures, training, reasoning-time interventions, embodied reasoning, and future directions (§C–I).

2 Foundations of Spatial Reasoning

♦ **Cognitive and Geometric Definitions.** Spatial reasoning involves representing and inferring

Proceedings of the 15th International Joint Conference on Natural Language Processing and the 5th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 1451–1479
November 6-10, 2026 ©2026 Association for Computational Linguistics

relationships among objects, encompassing both qualitative aspects like symbolic relations (containment, adjacency, direction) and quantitative aspects such as distance, angle, and scale (Tasnim et al., 2026; Chen et al., 2024). Classical frameworks distinguish topological relations invariant to deformation, projective relations dependent on viewpoint, and metric relations requiring explicit geometric grounding (Tasnim et al., 2026). While some topological and projective relations can be approximated from 2D cues, robust metric reasoning relies on depth, scale, or multi-view consistency (Pothiraj et al., 2025). These distinctions are crucial for evaluating foundation models, as competence in one type does not guarantee competence in others.

♦ **Spatial Reasoning vs. Visual Recognition.** Spatial reasoning differs from visual recognition, which focuses on identifying objects and attributes; spatial reasoning requires modeling relations and reasoning about configurations under transformation or partial views (Johnson et al., 2017; Hudson and Manning, 2019). Models often perform well on detection or captioning but fail basic relational or metric spatial queries (Suhr et al., 2017; Pothiraj et al., 2025), with dataset biases enabling shortcut learning of spatial correlations without true understanding (Wang and Ling, 2025). Recognition-based pipelines lack mechanisms for global spatial consistency, causing contradictions across viewpoints or counterfactuals (Jia et al., 2026a; Danish et al., 2025). Thus, spatial reasoning should be evaluated separately from perceptual performance.

♦ **Human Spatial Cognition and AI Parallels.** Human spatial cognition combines mental imagery, symbolic abstraction, and sensorimotor interaction to support mental rotation, perspective-taking, and navigation (Stogiannidis et al., 2025; Tasnim et al., 2026). These complex transformations reveal underlying spatial representations and impose cognitive load (McMains and Somers, 2005). Foundation models struggle especially when aligning egocentric and allocentric frames or with ambiguous cues (Pothiraj et al., 2025; Munje et al., 2025). Embodied navigation highlights the importance of memory, temporal integration, and action-conditioned perception, aspects static VLMs handle poorly (Song et al., 2025a; Zhang et al., 2025e). Advancing AI spatial reasoning likely requires architectures and training more closely aligned with human cognition. A glossary of spatial reasoning terms and technical terms for general understanding is provided in Table 2.

3 Spatial Reasoning in LLMs/VLMs

Here we examine how large language and vision–language models handle spatial reasoning across five complementary dimensions, highlighting key capabilities and limitations (Figure 2). Examples of each type are also available in Table 1.

Two dimensions deserve explicit clarification to avoid confusion. **Task-Level Reasoning** (§3.1) organizes work by *perceptual grounding and input modality*: text-only, image-based, video, 3D multi-view, and embodied. These categories differ in observability, interaction constraints, and sensory input. **Relation-Level Reasoning** (§3.2) instead organizes work by the *type of spatial relationship* being reasoned about: topological, directional, metric, perspective-dependent, and compositional. These relation types are orthogonal to task-level modality; for example, metric reasoning arises in both static image tasks and embodied navigation, while directional reasoning appears in text-only and 3D multi-view settings alike. Cross-references between these two dimensions are provided throughout.

3.1 Task-Level Reasoning

Spatial reasoning tasks for foundation models range from text-only linguistic inference to image, video, and spatiotemporal challenges that increase perceptual demands, often revealing reliance on 2D shortcuts and brittle reasoning (Li et al., 2024a; Sharma and Vats, 2025; Song et al., 2024; Xu et al., 2025b). The most complex tasks: 3D, multi-view, and embodied agentic settings, require viewpoint invariance, and long-term memory, showing that perceptual skills alone are insufficient without integrated depth reasoning and action-conditioned models.

■ **Text-Only Spatial Reasoning.** Text-only spatial tasks evaluate models' ability to manipulate spatial relations abstractly, isolating symbolic inference from perceptual noise, and include benchmarks for relative-position algebra, multi-step map reasoning, and hypothetical reconfiguration (Du et al., 2024; Li et al., 2024a; Liao et al., 2024; Ramakrishnan et al., 2025; Sharma and Vats, 2025). While LLMs often handle simple transitive or single-hop relations, they struggle with longer compositions and implicit metric constraints (Du et al., 2024; Li et al., 2024a; Sharma and Vats, 2025; Ramakrishnan et al., 2025). Several studies highlight that text-only tasks can overestimate real-world spatial competence due to the lack of grounding in depth or oc-

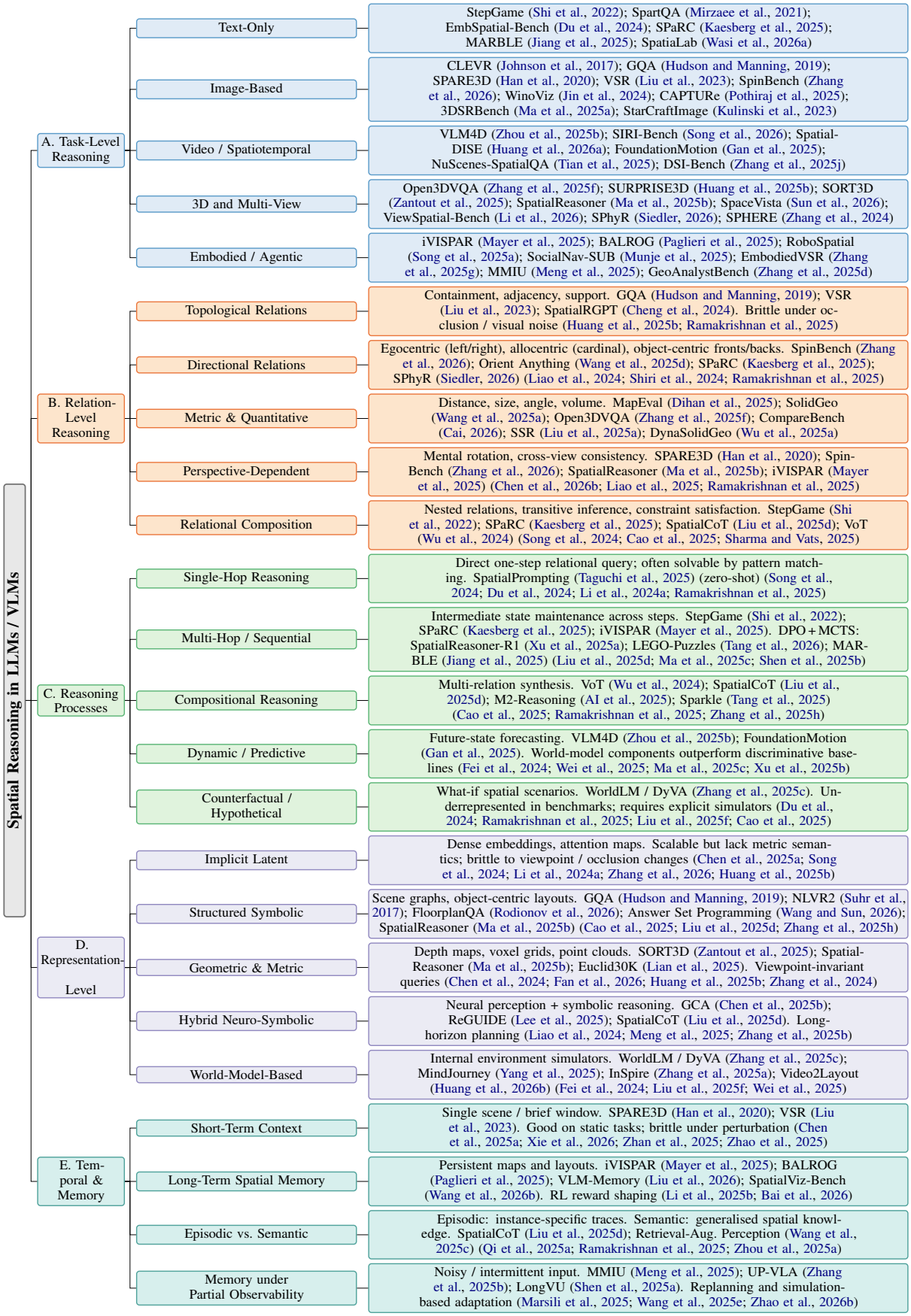


Figure 1: Multi-axis taxonomy of spatial reasoning in LLMs and VLMs. (A) Task-level input modality; (B) relation type; (C) reasoning process; (D) internal representation; (E) temporal and memory dimension. Leaf nodes list representative benchmarks and methods, with citations.

clusion, stressing the need for explicit mechanisms to map symbols to perceptual primitives (Song et al., 2024; Cao et al., 2025; Ma et al., 2025c). Datasets like StepGame and SpartQA exemplify this category, including more complex counterfactual and multi-hop variants that test compositional understanding but remain limited by the absence of visual grounding (Li et al., 2024b; Mirzaee et al., 2021; Mayer et al., 2025).

■ **Image-Based Spatial Reasoning.** Image-based reasoning on static scenes tests relational understanding of object position, occlusion, orientation, and geometry, dominating vision-language research (Song et al., 2024; Jin et al., 2024; Zhang et al., 2026, 2025h; Huang et al., 2025b; Han et al., 2020). Synthetic diagnostics (CLEVR, SPARE3D, VSR) and real-image benchmarks (GQA, NLVR2) show recognition or captioning success does not guarantee robust spatial reasoning, with persistent errors in distance, occlusion, orientation, and compositional relations caused by 2D shortcuts and dataset bias (Feng et al., 2026; Liao et al., 2024; Shiri et al., 2024; Zhang et al., 2026). Advances include richer supervision, scene-graph losses, relation-centric training, and planning-oriented models improving fine-grained and long-horizon inference, yet purely 2D supervision limits remain (Li et al., 2024d; Zhang et al., 2025h; Liu et al., 2025d; Zhang et al., 2025g). Challenging scenarios like multi-agent scenes, diagrams, multi-view consistency, and viewpoint-invariant 3D reasoning (e.g., StarCraftImage (Kulinski et al., 2023), 3DSR-Bench (Ma et al., 2025a)) highlight perspective-taking and cross-view alignment gaps, emphasizing the need for explicit relational and geometric modeling beyond static images (Sun et al., 2025; Xue et al., 2026). These limitations foreshadow the shortcut learning and metric deficiency failure modes detailed in §4.

■ **Video and Spatiotemporal Reasoning.** Video tasks require integrating temporal information, tracking objects, inferring motion and causality, and predicting future states, challenging frame-level models (Wei et al., 2025; Fei et al., 2024; Xu et al., 2025b; Wang et al., 2025d). Benchmarks like VLM4D, FoundationMotion, SIRI-Bench, SpatialDISE, and NuScenes-SpatialQA reveal naive frame aggregation fails to capture persistence, occlusion, and dynamics. Models with temporal state, memory, or dynamics modeling improve but still struggle with long-horizon generalization and annotation costs (Zhou et al., 2025b; Gan et al., 2025; Song et al., 2026; Huang et al., 2026a; Tian et al., 2025; Liu et al., 2025d). Recent methods explore world models, video diffusion reasoning, structured subspaces, and task-specific optimization to enhance dynamic spatial understanding, yet attention mechanisms often misalign with geometric features (Yang et al., 2025; Wang et al., 2025e; Ouyang, 2025). Spatiotemporal reasoning remains a frontier linking perception, dynamics, and planning, with current VLMs far from robust predictive and counterfactual reasoning (Ramakrishnan et al., 2025; Kamath et al., 2023; Li et al., 2025a).

■ **3D and Multi-View Spatial Reasoning.** 3D and multi-view tasks test volumetric understanding, viewpoint invariance, and metric consistency, critical for robotics and scene understanding but difficult for mostly 2D-trained models (Doh et al., 2025; Zhan et al., 2025; Zhang et al., 2024; Zhou et al., 2025a). Benchmarks like Open3DVQA (Zhang et al., 2025f), SURPRISE3D (Huang et al., 2025b), SORT3D (Zantout et al., 2025), and SpatialReasoner (Ma et al., 2025b) reveal failures in occlusion-aware reasoning, depth consistency, and multi-view feature alignment. Incorporating depth supervision, multi-view fusion, and geometric biases (e.g., point clouds, voxels) improves robustness but reduces scalability and label efficiency (Meng et al., 2024; Fan et al., 2026). Research advocates separating viewpoint alignment from recognition in evaluation and extends reasoning to generative and CAD tasks, where metric consistency governs 3D scene synthesis and design (Zhan et al., 2025; Zhang et al., 2026; Wang et al., 2025d; Fang, 2025; Ran et al., 2025; Wang et al., 2025b).

■ **Embodied and Agentic Spatial Reasoning.** Embodied spatial reasoning involves perception–action loops under partial observability, where planning, memory, affordance reasoning, and belief updating are as vital as perception. Navigation, manipulation, and interactive QA benchmarks reveal a knowing–doing gap: models can verbalize spatial facts but fail in consistent action, map maintenance, and error recovery, showing brittleness to embodiment shifts, sensor noise, and occlusion (Zhao et al., 2025; Zhang et al., 2025g; Liu et al., 2025f; Wang et al., 2025c; Ramakrishnan et al., 2025; Zhou et al., 2025a). Systems combining world models, episodic memory, symbolic planners, or reinforcement learning (RL) excel on long-horizon tasks, with evaluations emphasizing plan consistency and replanning over stepwise accuracy (Liu et al., 2025f; Zhao et al., 2025; Wang

et al., 2025c; Liu et al., 2025d; Zhao et al., 2026b). Recent work expands to bio-inspired navigation (BrainNav), program synthesis for 3D reasoning, language-grounded robotics, and efficient embodied models, showing diverse approaches and challenges in unifying evaluation across planning and control (Liao et al., 2024; Meng et al., 2025; Zhang et al., 2025b; Ishida, 2024; Bai et al., 2026; Marsili et al., 2025; Qi et al., 2025b; Mayer et al., 2025; Paglieri et al., 2025; Liu et al., 2025c; Traore et al., 2025).

These challenges directly instantiate the knowing–doing gap described in §4.

3.2 Relation-Level Reasoning

Relation-level reasoning focuses on how models understand and infer the *type* and properties of spatial relationships between objects, independent of the input modality (Task-Level, §3.1). These relation categories apply across text-only, image, video, 3D, and embodied settings. See Table 1 for examples.

■ **Topological Relations.** Topological relations like containment, adjacency, and support are qualitative and invariant to deformation, making them among the easiest spatial relations for LLMs and VLMs, with good performance on benchmarks such as GQA and VSR, though often relying on dataset regularities rather than true relational abstraction (Song et al., 2024; Du et al., 2024; Jin et al., 2024; Zhang et al., 2026, 2025h; Cheng et al., 2024; Hudson and Manning, 2019; Liu et al., 2023). However, accuracy declines sharply in compositional contexts, under occlusion, or with visual noise, showing that simple predicate competence is necessary but insufficient for robust reasoning (Huang et al., 2025b; Liao et al., 2024; Ramakrishnan et al., 2025; Mayer et al., 2025). Therefore, targeted diagnostics and counterfactual perturbations remain crucial to separate superficial pattern learning from genuine topological understanding (Song et al., 2024; Ramakrishnan et al., 2025; Sharma and Vats, 2025).

■ **Directional Relations.** Directional relations like egocentric (left/right), allocentric (cardinal directions), and object-centric fronts/backs require clear reference frames and are sensitive to viewpoint shifts and annotation differences, causing systematic errors when frames or object axes misalign with camera axes (Wang et al., 2025d; Jin et al., 2024; Zhang et al., 2025h; Liao et al., 2024). While adding orientation supervision or explicit frame encoding improves robustness, no standard diagnostics for frame sensitivity exist, making directional reasoning a persistent challenge across datasets and real-world versus simulation settings (Wang et al., 2025d; Zhang et al., 2026; Cao et al., 2025; Liu et al., 2025d). Empirical studies show LLMs handle topological relations better but struggle with directional and metric ones, especially in multi-hop and dynamic composition tasks (Rizvi et al., 2024; Liao et al., 2024; Meng et al., 2025; Zhang et al., 2025b; Song et al., 2026), while 3D simulations and benchmarks like (Siedler, 2026) highlight ongoing difficulties with qualitative spatial relations and orientation consistency (Li et al., 2024c; Gholami et al., 2025; Han et al., 2020; Liu et al., 2023; Zantout et al., 2025). Overall, viewpoint-dependence and metric inconsistency make directional reasoning a major failure mode in both static and video spatial data (Shiri et al., 2024; Ramakrishnan et al., 2025).

■ **Metric and Quantitative Relations.** Metric relations involving distances, sizes, angles, and volumes reveal a key limitation in many foundation models: lacking stable internal scale or depth awareness, they often confuse relative scale cues with dataset biases, leading to systematic errors in scenes with unusual camera settings or object sizes (Zhang et al., 2025f; Huang et al., 2025b; Feng et al., 2026; Zhang et al., 2024; Liao et al., 2025). Incorporating explicit geometric components like depth prediction, point-cloud reasoning, or metric losses improves performance but incurs annotation and computational costs and can reduce robustness to domain shifts (Zhang et al., 2025f; Cao et al., 2025; Zhang et al., 2024; Liao et al., 2025). Consequently, metric reasoning requires architectural grounding or targeted supervision rather than arising from large-scale pretraining alone (Huang et al., 2025b; Zhao et al., 2026a; Zhang et al., 2025i). Geo-spatial benchmarks such as MapEval (Dihan et al., 2025) and others like SolidGeo (Wang et al., 2025a), CompareBench (Cai, 2026), DynaSolidGeo (Wu et al., 2025a), and SSR (Liu et al., 2025a) demonstrate persistent deficiencies in metric understanding compared to humans. This persistent deficiency corresponds to the metric failure mode in §4 and motivates the geometric and metric representations discussed in §3.1.

■ **Perspective-Dependent Relations.** Relations requiring mental rotation, mirror/reflection reasoning, and cross-perspective consistency demand internal transformations or viewpoint normalization,

but VLMs and LLMs commonly fail on rotated or mirrored scenes unless trained with multi-view data or equipped with explicit transformation operators (Zhang et al., 2025d; Jin et al., 2024; Wang et al., 2025d; Zhang et al., 2026, 2024; Liu et al., 2025d). While augmenting training with synthetic view perturbations, contrastive multi-view objectives, or geometric priors improves robustness (Zhan et al., 2025; Huang et al., 2025b; Liao et al., 2025), compositional tasks requiring chained transformations remain challenging, underscoring that viewpoint invariance depends on appropriate inductive biases and representations, not just more data (Jin et al., 2024; Wang et al., 2025d; Ramakrishnan et al., 2025). Perspective-dependent relations thus expose a fundamental limitation in current models' 3D spatial simulation abilities (Chen et al., 2026b; Ma et al., 2025b), with tasks like iVISPAR's Sliding Geom Puzzle explicitly testing these cognitive skills (Mayer et al., 2025).

■ **Relational Composition.** Relational composition, which involves nested relations, transitive inference, and constraint satisfaction, tests whether models can combine primitives into complex explanations, but benchmarks requiring nested reasoning, such as *A is left of B, B is in front of C, where is A relative to C?*, show many models fail despite handling primitives well (Song et al., 2024; Du et al., 2024; Sharma and Vats, 2025; Zhang et al., 2026; Ramakrishnan et al., 2025). Research indicates that structured representations like scene graphs and object-centric slots, alongside explicit multi-step inference methods such as chain-of-thought and spatial COT, improve compositional generalization but introduce trade-offs in speed and complexity (Cao et al., 2025; Liu et al., 2025d; Zhang et al., 2025h; Ma et al., 2025c). Relational composition remains a challenging benchmark for robust spatial inference, with tasks like StepGame and SPaRC highlighting current limits of chain-of-thought and tree-of-thought prompting (Song et al., 2024; Ramakrishnan et al., 2025; Sharma and Vats, 2025; Li et al., 2024b; Kaesberg et al., 2025; Wu et al., 2024).

3.3 Reasoning-Processes

Here we explore the underlying cognitive and computational processes that enable spatial reasoning in LLMs and VLMs:

■ **Single-Hop Reasoning.** Single-hop reasoning involves direct one-step relational queries often solvable by pattern matching or associative retrieval, which many models excel at through local attention or nearest-neighbor search over embeddings rather than true reasoning chains (Song et al., 2024; Du et al., 2024; Li et al., 2024a; Sharma and Vats, 2025; Ramakrishnan et al., 2025). This success risks overestimating spatial competence, especially when benchmarks are dominated by single-hop items, motivating diagnostic splits and adversarial examples to expose shortcut strategies (Song et al., 2024; Zhang et al., 2026; Ramakrishnan et al., 2025), and highlighting that single-hop ability is necessary but insufficient for compositional or dynamic tasks (Du et al., 2024; Sharma and Vats, 2025; Cao et al., 2025). Prompting methods like SpatialPrompting can elicit zero-shot spatial reasoning from multimodal models using keyframe-driven prompts, but these gains are fragile and sensitive to prompt design, indicating limited internalization of spatial concepts (Taguchi et al., 2025; Ye et al., 2025; Liu et al., 2023).

■ **Multi-Hop and Sequential Reasoning.** Multi-hop and sequential spatial reasoning require maintaining intermediate states across steps, with benchmarks like StepGame (Li et al., 2024b), SPaRC (Kaesberg et al., 2025), and iVISPAR (Mayer et al., 2025) revealing sharp performance drops as hop depth increases, indicating many models lack mechanisms to reliably update memory and representations (Sharma and Vats, 2025; Song et al., 2024; Ramakrishnan et al., 2025; Xu et al., 2025b; Liu et al., 2025d). Approaches using explicit intermediate structures like scene graphs, slot-based memory, and chain-of-thought prompting improve robustness but face challenges with error accumulation (Liu et al., 2025d; Ma et al., 2025c; Cao et al., 2025), while optimization methods such as SpatialReasoner-R1's Direct Preference Optimization with Monte Carlo Tree Search have further advanced complex multi-step reasoning (Xu et al., 2025a; Shen et al., 2025b). This capacity is essential for spatial planning and navigation, with ongoing research highlighting gaps through datasets like LegoPuzzles and MARBLE (Liao et al., 2024; Meng et al., 2025; Song et al., 2026; Tang et al., 2026; Shi et al., 2022; Jiang et al., 2025) and RL approaches pushing multi-tool coordination (Pan and Liu, 2025; Li et al., 2025b; Chen et al., 2026a).

■ **Compositional Reasoning.** Compositional reasoning synthesizes multiple spatial relations and object attributes, with approaches like Visualization-of-Thought prompting (Wu et al., 2024) showing some gains though generalization remains limited.

Studies find structured symbolic pipelines often outperform end-to-end embeddings in systematicity and interpretability for complex compositions (Cao et al., 2025; Liu et al., 2025d; Zhang et al., 2025h; Ramakrishnan et al., 2025), while spatial chain-of-thought prompting helps but is sensitive to design and distribution shifts (Liu et al., 2025d; Sharma and Vats, 2025; Ma et al., 2025c). Unified frameworks like M2-Reasoning improve compositional inference through multi-task training (AI et al., 2025), and curriculum methods use stress mastering spatial primitives to boost generalization on multi-relation tasks (Tang et al., 2025).

■ **Dynamic and Predictive Reasoning.** Dynamic and predictive reasoning focuses on forecasting future spatial and temporal states, with studies like VLM4D (Zhou et al., 2025b) and FoundationMotion (Gan et al., 2025) showing that static aggregation limits true predictive capacity. Models incorporating latent dynamics or explicit world-model components outperform discriminative baselines in video reasoning and agentic tasks (Fei et al., 2024; Wei et al., 2025; Ma et al., 2025c; Xu et al., 2025b), but building accurate, stable dynamics models remains challenging due to drift and expressivity trade-offs (Fei et al., 2024; Liu et al., 2025f). Evaluation benefits from probabilistic metrics and counterfactual tests (Ma et al., 2025c; Ramakrishnan et al., 2025), while video benchmarks and datasets (Zhou et al., 2025c; Huang et al., 2025a; Gan et al., 2025) advance spatiotemporal memory integration and dynamic reasoning including rotations, translations, and cognitive map reconstruction.

■ **Counterfactual and Hypothetical Reasoning.** Counterfactual and hypothetical reasoning involves *what-if* spatial scenarios requiring explicit world-model-based representations (Zhang et al., 2025c) and remains a significant gap in current VLMs and LLMs. These tasks test mental simulation and causal inference about transformed objects but are underrepresented in benchmarks, with early studies showing models trained only on observational data struggle with consistent counterfactual predictions (Du et al., 2024; Ramakrishnan et al., 2025; Ma et al., 2025c). Promising methods combine explicit simulators, differentiable physics, and intervention-based augmentation (Liu et al., 2025f; Fei et al., 2024; Cao et al., 2025), yet scaling these approaches and establishing standardized benchmarks remain critical challenges (Ma et al., 2025c; Sharma and Vats, 2025).

3.4 Reasoning at Representation-Level

Here we review key approaches to spatial reasoning at the representation level:

■ **Implicit Latent Representations.** Implicit latent representations, such as dense embeddings and attention maps, dominate end-to-end VLMs due to their scalability but suffer from poor interpretability and generalization in spatial tasks (Mayer et al., 2025; Wang et al., 2026a). While these latent encodings capture useful statistical regularities, they often lack stable metric semantics and fail under changes in viewpoint, scale, or occlusion (Song et al., 2024; Li et al., 2024a; Liu et al., 2025d; Zhang et al., 2026; Huang et al., 2025b). Their inability to explicitly represent objects or relations hampers reliable compositional and counterfactual reasoning, motivating hybrid methods that combine latent efficiency with structured inductive biases (Cao et al., 2025; Liu et al., 2025d; Ramakrishnan et al., 2025). Consequently, implicit encodings remain a practical choice but require complementary mechanisms for advanced spatial reasoning (Song et al., 2024; Li et al., 2024a; Liu et al., 2025d).

■ **Structured Symbolic Representations.** Structured symbolic representations like scene graphs and object-centric layouts, used in benchmarks such as GQA (Hudson and Manning, 2019) and NLVR2 (Sarkar et al., 2026), improve compositionality and interpretability by explicitly exposing objects and predicates, supporting robust generalization on nested and multi-hop tasks (Rodionov et al., 2026; Cao et al., 2025; Liu et al., 2025d; Zhang et al., 2025h). However, they depend on accurate object detection and relation extraction, with perception errors cascading through symbolic layers and increased dataset costs (Cao et al., 2025; Zhang et al., 2025h). Hybrid approaches combining neural perception with symbolic reasoning cores balance scalability and systematicity (Cao et al., 2025; Liu et al., 2025d; Ramakrishnan et al., 2025), with recent advances including 3D point cloud fusion (Jahangard et al., 2025), Answer Set Programming pipelines (Wang and Sun, 2026), and SpatialReasoner frameworks integrating LLM spatial reasoning and explicit visual properties for 3D grounding (Liu et al., 2025e; Premsri and Kordjamshidi, 2025).

■ **Geometric and Metric Representations.** Geometric and metric representations, such as: depth maps, voxel grids, and point clouds, are vital for robotics and navigation, with models like Spatial-

Reasoner (Ma et al., 2025b) and SORT3D (Zantout et al., 2025) leveraging explicit 3D data for better spatial manipulation (Han et al., 2026). Techniques such as multi-view aggregation and depth supervision improve tasks like distance estimation and occlusion handling despite annotation and domain challenges (Zhang et al., 2025f; Huang et al., 2025b; Cao et al., 2025). These representations enable viewpoint-invariant queries but need precise calibration in embodied settings (Zhang et al., 2024; Zhou et al., 2025a), making geometric methods preferred for precision tasks and hybrids suited for broader perception (Cao et al., 2025; Huang et al., 2025b; Zhang et al., 2025i). Geometry-focused supervision like Euclid30K further enhances metric reasoning via fine-tuning on surrogate tasks (Lian et al., 2025).

■ **Hybrid Neuro-Symbolic Representations.** Hybrid neuro-symbolic methods merge neural perception with symbolic reasoning, leveraging flexibility alongside modular, interpretable systems (Cao et al., 2025; Liu et al., 2025d; Zhang et al., 2025h; Ramakrishnan et al., 2025; Liu et al., 2025f). For example, ReGUIDE (Lee et al., 2025) grounds GUI elements via spatial criticism and search, showcasing this promising direction. These hybrids excel at explicit object manipulation, logical consistency, and structured reasoning in dynamic contexts (Liu et al., 2025d; Zhang et al., 2025g; Cao et al., 2025). Recent work integrates neural embeddings with symbolic spatial planners and scene graphs, enabling compositional reasoning and long-horizon planning, overcoming latent model limitations (Liao et al., 2024; Meng et al., 2025; Zhang et al., 2025b). Notably, the Geometrically-Constrained Agent (GCA) improves robustness by decoupling semantic perception from geometric constraint satisfaction during execution (Chen et al., 2025b).

■ **World-Model-Based Representations.** World-model-based representations, like internal environment simulators in WorldLM and DyVA (Zhang et al., 2025c), are crucial for dynamic spatial understanding and planning but remain underdeveloped. Learned latent dynamics models enable planning, imagination, and counterfactual reasoning by rolling forward explicit states under actions, supporting embodied tasks despite challenges with high-dimensional or partial observations (Fei et al., 2024; Liu et al., 2025f; Wei et al., 2025; Ma et al., 2025c), while requiring alignment of latent states with interpretable geometric variables for precise metric reasoning (Cao et al., 2025). Recent progress includes video diffusion world models for test-time reasoning (Yang et al., 2025), intrinsic spatial reasoning in vision-language-action robotics (Zhang et al., 2025a), and cognitive map reconstruction that supports spatial memory and planning (Huang et al., 2026b).

3.5 Temporal and Memory Dimensions

Here we review spatial reasoning across different temporal and memory contexts:

■ **Short-Term Spatial Context.** Short-term spatial context involves reasoning within a single scene or brief temporal window, focusing on local object relations and precise spatial layout, as benchmarked by SPARE3D and VSR (Han et al., 2020; Liu et al., 2023). Many vision-language models perform well here by leveraging powerful visual encoders and implicit attention (Chen et al., 2025a; Xie et al., 2026; Fei et al., 2024; Liu et al., 2025f), but this success often masks brittleness to viewpoint changes, occlusion, or temporal perturbations (Zhan et al., 2025; Zhao et al., 2025). Consequently, short-term spatial reasoning does not reliably generalize to tasks requiring memory or planning beyond the immediate scene.

■ **Long-Term Spatial Memory (Maps and Layouts).** Long-term spatial memory maintains persistent maps or layouts across extended time horizons, enabling navigation and multi-step planning (Mayer et al., 2025; Paglieri et al., 2025). Explicit memory modules and learned spatial maps improve long-horizon reasoning in embodied and video tasks (Liu et al., 2026; Bai et al., 2026), yet benchmarks reveal a large gap between models relying on short-term perception and those with spatial memory retention (Wang et al., 2026b). RL studies show that shaping rewards around memory use enhances planning and consistency, though maintaining metric and topological integrity remains challenging under noise (Li et al., 2025b).

■ **Episodic vs. Semantic Spatial Memory.** Spatial reasoning often distinguishes episodic memory, which stores recent instance-specific traces, from semantic memory capturing generalized spatial knowledge (Qi et al., 2025a; Liu et al., 2025b). Models integrating both types outperform those relying on only one, especially in repeated exploration and instruction-following under uncertainty (Liu et al., 2025d; Wang et al., 2025c; Zhou et al., 2025a). Episodic memory is critical for maintaining temporal coherence and resolving occlusion or

viewpoint ambiguity, but balancing episodic specificity and semantic abstraction remains a difficult open problem (Ramakrishnan et al., 2025; Mayer et al., 2025).

■ **Memory Updating under Partial Observability.** Updating spatial memory with incomplete, noisy, or intermittent inputs demands integrating new evidence while avoiding catastrophic forgetting or drift (Wang et al., 2025c). Hybrid approaches combining learned updates with structural constraints show improved robustness in navigation and video reasoning (Meng et al., 2025; Zhang et al., 2025b), and temporally grounded methods emphasize memory for motion prediction and event sequencing (Huang et al., 2025a; Shen et al., 2025a). Agentic strategies like replanning and simulation-based adaptation further support dynamic memory revision in evolving environments (Marsili et al., 2025; Wang et al., 2025e), while RL benchmarks test spatial grounding through interaction and feedback (Zhao et al., 2026b).

4 Failure Modes and Open Problems

Spatial reasoning in foundation models exhibits four recurrent failure modes. For each, we identify which methodological approaches partially mitigate it and where gaps remain (Pothiraj et al., 2025; Wang and Ling, 2025).

Shortcut learning on 2D datasets. Models exploit superficial appearance correlations rather than genuine relational structure, yielding poor generalization under viewpoint changes or counterfactual perturbations (Johnson et al., 2017; Wang and Ling, 2025). Counterfactual and diagnostic evaluation (e.g., synthetic diagnostics, relational swaps) exposes this failure; structured symbolic representations and scene-graph losses partially mitigate it but do not eliminate it (Suhr et al., 2017; Pothiraj et al., 2025).

Metric reasoning deficiency. Models fail to maintain consistent internal scale even with depth or 3D supervision, causing compounding errors in navigation and manipulation (Chen et al., 2024; Pothiraj et al., 2025). Geometric and metric representations (depth maps, point clouds, metric supervision) improve distance estimation but are costly and do not generalize robustly across domains (Danish et al., 2025; Wang and Ling, 2025).

Perspective and rotation errors. Models confuse egocentric, allocentric, and object-centric frames, producing inconsistent answers across viewpoints (Stogiannidis et al., 2025; Danish et al., 2025). Multi-view contrastive training and geometry-aware encodings yield partial improvement; chain-of-thought prompting provides limited and fragile gains (Jia et al., 2026a; Wang and Ling, 2025).

The knowing–doing gap. Agents verbalize correct spatial relations but execute spatially incoherent actions, reflecting a disconnect between declarative knowledge and procedural control (Song et al., 2025a). Vision–language–action models and RL with spatially shaped rewards reduce but do not close this gap (Zhang et al., 2025e; Pothiraj et al., 2025).

Table 3 in Appendix H provides a structured mapping of each failure mode to the representation and training strategies that address it and the conditions under which those strategies still fail.

Three cross-cutting patterns emerge. First, **scaling shows diminishing returns**: larger models improve simple topological and directional tasks, but gains plateau on metric, perspective-dependent, and compositional reasoning (Wasi et al., 2026b; Jia et al., 2026b). Second, **chain-of-thought prompting yields brittle improvements**: structured decomposition helps multi-hop reasoning but fails under distribution shifts and prompt changes, revealing weak underlying representations (Liu et al., 2025d; Jia et al., 2026a). Third, **encoder choice alone is insufficient**: although Vision Transformers improve global relational context over CNNs, persistent metric and perspective errors suggest deeper reasoning-stack limitations (Johnson et al., 2017; Pothiraj et al., 2025).

5 Conclusion

Spatial reasoning in LLMs/VLMs remains fragmented, as models consistently struggle with metric reasoning, perspective alignment, long-term memory, and action grounding across static, 3D, and embodied settings. Scaling alone is insufficient, as many improvements depend on prompting strategies or external tools instead of coherent spatial representations. We argue for architectural shifts toward structured world models with targeted supervision for prediction, counterfactual reasoning, and planning. Key priorities include benchmarks with counterfactual perturbations and metric consistency checks, training pipelines coupling metric supervision with reasoning objectives, and evaluation within full perception-action loops.

Limitations

This survey focuses primarily on spatial reasoning within large language and vision-language models, which may not fully capture challenges in other AI modalities or specialized domains. Our analysis relies on existing benchmarks and literature, which themselves exhibit biases and uneven coverage of spatial phenomena. Additionally, rapid advancements in model architectures and datasets mean some insights may evolve quickly.

References

Inclusion AI, Fudong Wang, Jiajia Liu, Jingdong Chen, Jun Zhou, Kaixiang Ji, Lixiang Ru, Qingpei Guo, Ruobing Zheng, Tianqi Li, Yi Yuan, Yifan Mao, Yuting Xiao, and Ziping Ma. 2025. M2-reasoning: Empowering mllms with unified general and spatial reasoning. *Preprint*, arXiv:2507.08306.

Jiayang Ao, Yanbei Jiang, Qiuhong Ke, and Krista A. Ehinger. 2025. Open-world amodal appearance completion. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6490–6499.

Qianqian Bai, Zhongpu Chen, Ling Luo, Huaming Du, Yuqian Lei, and Ziyun Jiao. 2026. Endowing embodied agents with spatial reasoning capabilities for vision-and-language navigation. *Preprint*, arXiv:2504.08806.

Jie Cai. 2026. Comparebench: A benchmark for visual comparison reasoning in vision-language models. *Preprint*, arXiv:2509.22737.

Bing Cao, Quanhao Lu, Jiekang Feng, Qilong Wang, Qinghua Hu, and Pengfei Zhu. 2025. Efficient masked autoencoder for video object counting and a large-scale benchmark. *Preprint*, arXiv:2411.13056.

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

Shiqi Chen, Tongyao Zhu, Ruochen Zhou, Jinghan Zhang, Siyang Gao, Juan Carlos Niebles, Mor Geva, Junxian He, Jiajun Wu, and Manling Li. 2025a. Why is spatial reasoning hard for vlms? an attention mechanism perspective on focus areas. *Preprint*, arXiv:2503.01773.

Siyi Chen, Mikaela Angelina Uy, Chan Hee Song, Faisal Ladhak, Adithyavairavan Murali, Qing Qu, Stan Birchfield, Valts Blukis, and Jonathan Tremblay. 2026a. Spacetools: Tool-augmented spatial reasoning via double interactive rl. *Preprint*, arXiv:2512.04069.

Zeren Chen, Xiaoya Lu, Zhijie Zheng, Pengrui Li, Lehan He, Yijin Zhou, Jing Shao, Bohan Zhuang, and Lu Sheng. 2025b. Geometrically-constrained agent for spatial reasoning. *Preprint*, arXiv:2511.22659.

Zhangquan Chen, Manyuan Zhang, Xinlei Yu, Xufang Luo, Mingze Sun, Zihao Pan, Xiang An, Yan Feng, Peng Pei, Xunliang Cai, and Ruqi Huang. 2026b. Think with 3d: Geometric imagination grounded spatial reasoning from limited views. *Preprint*, arXiv:2510.18632.

An-Chieh Cheng, Hongxu Yin, Yang Fu, Qiushan Guo, Ruihan Yang, Jan Kautz, Xiaolong Wang, and Sifei Liu. 2024. SpatialRGPT: Grounded spatial reasoning in vision-language models. In *Advances in Neural Information Processing Systems 37*, Advances in Neural Information Processing Systems, pages 135062–135093. Neural Information Processing Systems Foundation, Inc. (NeurIPS).

Muhammad Sohail Danish, Muhammad Akhtar Munir, Syed Roshaan Ali Shah, Kartik Kuckreja, Fahad Shahbaz Khan, Paolo Fraccaro, Alexandre Lacoste, and Salman Khan. 2025. Geobench-vlm: Benchmarking vision-language models for geospatial tasks. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7132–7142.

Nianchen Deng, Lixin Gu, Shenglong Ye, Yinan He, Zhe Chen, Songze Li, Haomin Wang, Xingguang Wei, Tianshuo Yang, Min Dou, Tong He, Wenqi Shao, Kaipeng Zhang, Yi Wang, Botian Shi, Yanting Zhang, Jifeng Dai, Yu Qiao, Hongjie Zhang, and Wenhai Wang. 2025. Internspatial: A comprehensive dataset for spatial reasoning in vision-language models. *Preprint*, arXiv:2506.18385.

Mahir Labib Dihan, Md Tanvir Hassan, Md Tanvir Parvez, Md Hasebul Hasan, Md Almash Alam, Muhammad Aamir Cheema, Mohammed Eunus Ali, and Md Rizwan Parvez. 2025. Mapeval: A map-based evaluation of geo-spatial reasoning in foundation models. *Preprint*, arXiv:2501.00316.

Hyungjun Doh, Dong In Lee, Seunggeun Chi, Pin-Hao Huang, Kwonjoon Lee, Sangpil Kim, and Karthik Ramani. 2025. Occlusion-aware temporally consistent amodal completion for 3d human-object interaction reconstruction. In *Proceedings of the 33rd ACM International Conference on Multimedia*, MM '25, page 52–61, New York, NY, USA. Association for Computing Machinery.

Mengfei Du, Binhao Wu, Zejun Li, Xuanjing Huang, and Zhongyu Wei. 2024. EmbSpatial-bench: Benchmarking spatial understanding for embodied tasks with large vision-language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 346–355, Bangkok, Thailand. Association for Computational Linguistics.

Zhiwen Fan, Jian Zhang, Renjie Li, Junge Zhang, Runjin Chen, Hezhen Hu, Kevin Wang, Huaizhi Qu,

Shijie Zhou, Dilin Wang, Zhicheng Yan, Hongyu Xu, Justin Theiss, Tianlong Chen, Jiachen Li, Zhengzhong Tu, Zhangyang Wang, and Rakesh Ranjan. 2026. Vlm-3r: Vision-language models augmented with instruction-aligned 3d reconstruction. *Preprint*, arXiv:2505.20279.

Jiading Fang. 2025. Embodied spatial intelligence: from implicit scene modeling to spatial reasoning. *Preprint*, arXiv:2509.00465.

Hao Fei, Shengqiong Wu, Wei Ji, Hanwang Zhang, Meishan Zhang, Mong-Li Lee, and Wynne Hsu. 2024. Video-of-thought: Step-by-step video reasoning from perception to cognition. In *Forty-first International Conference on Machine Learning*.

Jie Feng, Jinwei Zeng, Qingyue Long, Hongyi Chen, Jie Zhao, Yanxin Xi, Zhilun Zhou, Yuan Yuan, Shengyuan Wang, Qingbin Zeng, Songwei Li, Yunke Zhang, Yuming Lin, Tong Li, Jingtao Ding, Chen Gao, Fengli Xu, and Yong Li. 2025. Large language models for robot learning: A survey. *arXiv preprint*.

Zhiyuan Feng, Zhaolu Kang, Qijie Wang, Zhiying Du, Jiongrui Yan, Shubin Shi, Chengbo Yuan, Huizhi Liang, Yu Deng, Qixiu Li, Rushuai Yang, Arctanx An, Leqi Zheng, Weijie Wang, Shawn Chen, Sicheng Xu, Yaobo Liang, Jiaolong Yang, and Baining Guo. 2026. Seeing across views: Benchmarking spatial reasoning of vision-language models in robotic scenes. *Preprint*, arXiv:2510.19400.

Yulu Gan, Ligeng Zhu, Dandan Shan, Baifeng Shi, Hongxu Yin, Boris Ivanovic, Song Han, Trevor Darrell, Jitendra Malik, Marco Pavone, and Boyi Li. 2025. Foundationmotion: Auto-labeling and reasoning about spatial movement in videos. *Preprint*, arXiv:2512.10927.

Mohsen Gholami, Ahmad Rezaei, Zhou Weimin, Sitong Mao, Shunbo Zhou, Yong Zhang, and Mohammad Akbari. 2025. Spatial reasoning with vision-language models in ego-centric multi-view scenes. *Preprint*, arXiv:2509.06266.

Wenyu Han, Siyuan Xiang, Chenhui Liu, Ruoyu Wang, and Chen Feng. 2020. Spare3d: A dataset for spatial reasoning on three-view line drawings. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14678–14687.

Yi Han, Enshen Zhou, Shanyu Rong, Jingkun An, Pengwei Wang, Zhongyuan Wang, Cheng Chi, Lu Sheng, and Shanghang Zhang. 2026. Tiger: Tool-integrated geometric reasoning in vision-language models for robotics. *Preprint*, arXiv:2510.07181.

Haojian Huang, Haodong Chen, Shengqiong Wu, Meng Luo, Jinlan Fu, Xinya Du, Hanwang Zhang, and Hao Fei. 2025a. $\mathcal{V}ista\mathcal{DPO}$: Video hierarchical spatial-temporal direct preference optimization for large video models. In *Forty-second International Conference on Machine Learning*.

Jiaxin Huang, Ziwen Li, Hanlue Zhang, Runnan Chen, Zhengqing Gao, Xiao He, Yandong Guo, Wenping Wang, Tongliang Liu, and Mingming Gong. 2025b. SURPRISE3D: A dataset for spatial understanding and reasoning in complex 3d scenes. In *Advances in Neural Information Processing Systems 38*, Advances in Neural Information Processing Systems, pages 127897–127936. Neural Information Processing Systems Foundation, Inc. (NeurIPS).

Xinmiao Huang, Qisong He, Zhenglin Huang, Boxuan Wang, Zhuoyun Li, Guangliang Cheng, Yi Dong, and Xiaowei Huang. 2026a. Spatial-dise: A unified benchmark for evaluating spatial reasoning in vision-language models. *Preprint*, arXiv:2510.13394.

Yibin Huang, Wang Xu, Wanyue Zhang, Helu Zhi, Jingjing Huang, Yangbin Xu, Yangang Sun, Conghui Zhu, and Tiejun Zhao. 2026b. Video2layout: Recall and reconstruct metric-grounded cognitive map for spatial reasoning. *Preprint*, arXiv:2511.16160.

Drew A. Hudson and Christopher D. Manning. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6693–6702.

Shu Ishida. 2024. *Spatial reasoning and planning for deep embodied agents*. Ph.D. thesis, University of Oxford.

Amirmohammad Izadi, Mohammadali Banayeeanzade, Fatemeh Askari, Ali Rahimiakbar, Mohammad Vahedi, Hosein Hasani, and Mahdieh Baghshah. 2025. Visual structures helps visual reasoning: Addressing the binding problem in VLMs. In *Advances in Neural Information Processing Systems 38*, Advances in Neural Information Processing Systems, pages 180813–180851. Neural Information Processing Systems Foundation, Inc. (NeurIPS).

Simindokht Jahangard, Mehrzad Mohammadi, Abhinav Dhall, and Hamid Rezatofighi. 2025. A multi-modal neuro-symbolic approach for spatial reasoning-based visual grounding in robotics. *Preprint*, arXiv:2510.27033.

Binbin Ji, Siddharth Agrawal, Qiance Tang, and Yvonne Wu. 2025. Enhancing spatial reasoning in vision-language models via chain-of-thought prompting and reinforcement learning. *Preprint*, arXiv:2507.13362.

Mengdi Jia, Zekun Qi, Shaochen Zhang, Wenyao Zhang, Xinqiang Yu, Jiawei He, He Wang, and Li Yi. 2026a. Omnispatial: Towards comprehensive spatial reasoning benchmark for vision language models. *Preprint*, arXiv:2506.03135.

Mengdi Jia, Zekun Qi, Shaochen Zhang, Wenyao Zhang, XinQiang Yu, Jiawei He, He Wang, and Li Yi. 2026b. Omnispatial: Towards comprehensive spatial reasoning benchmark for vision language models. In *The Fourteenth International Conference on Learning Representations*.

Yulun Jiang, Yekun Chai, Maria Brbić, and Michael Moor. 2025. Marble: A hard benchmark for multimodal spatial reasoning and planning. *Preprint*, arXiv:2506.22992.

Woojeong Jin, Tejas Srinivasan, Jesse Thomason, and Xiang Ren. 2024. WINOVIZ: Probing visual properties of objects under different states. In *Proceedings of the Fifth Workshop on Insights from Negative Results in NLP*, pages 110–123, Mexico City, Mexico. Association for Computational Linguistics.

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross Girshick. 2017. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1988–1997.

Lars Benedikt Kaesberg, Jan Philip Wahle, Terry Ruas, and Bela Gipp. 2025. SPaRC: A spatial pathfinding reasoning challenge. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 10359–10390, Suzhou, China. Association for Computational Linguistics.

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

Sean Kulinski, Nicholas R. Waytowich, James Z. Hare, and David I. Inouye. 2023. Starcraftimage: A dataset for prototyping spatial reasoning methods for multi-agent environments. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22004–22013.

Hyunseok Lee, Jeonghoon Kim, Beomjun Kim, Jihoon Tack, Chansong Jo, Jaehong Lee, Cheonbok Park, Sookyo In, Jinwoo Shin, and Kang Min Yoo. 2025. Reguide: Data efficient gui grounding via spatial reasoning and search. *Preprint*, arXiv:2505.15259.

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

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. 2026. Viewspatial-bench: Evaluating multi-perspective spatial localization in vision-language models. In *Computer Vision – ECCV 2026*, pages 95–111, Cham. Springer Nature Switzerland.

Fangjun Li, David C. Hogg, and Anthony G. Cohn. 2024b. 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. AAAI Press.

Fangjun Li, David C. Hogg, and Anthony G. Cohn. 2024c. Reframing spatial reasoning evaluation in language models: A real-world simulation benchmark for qualitative reasoning. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*, pages 6342–6349. International Joint Conferences on Artificial Intelligence Organization. Main Track.

Rongjie Li, Songyang Zhang, Dahua Lin, Kai Chen, and Xuming He. 2024d. From pixels to graphs: Open-vocabulary scene graph generation with vision-language models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 28076–28086.

Yifan Li, Zhixin Lai, Wentao Bao, Zhen Tan, Anh Dao, Kewei Sui, Jiayi Shen, Dong Liu, Huan Liu, and Yu Kong. 2025a. Visual large language models for generalized and specialized applications. *Preprint*, arXiv:2501.02765.

Zongzhao Li, Zongyang Ma, Mingze Li, Songyou Li, Yu Rong, Tingyang Xu, Ziqi Zhang, Deli Zhao, and Wenbing Huang. 2025b. Star-r1: Spatial transformation reasoning by reinforcing multimodal llms. *Preprint*, arXiv:2505.15804.

Shijie Lian, Changti Wu, Laurence Tianruo Yang, Hang Yuan, Bin Yu, Lei Zhang, and Kai Chen. 2025. Euclid's gift: Enhancing spatial perception and reasoning in vision-language models via geometric surrogate tasks. *Preprint*, arXiv:2509.24473.

Yuan-Hong Liao, Rafid Mahmood, Sanja Fidler, and David Acuna. 2024. Reasoning paths with reference objects elicit quantitative spatial reasoning in large vision-language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17028–17047, Miami, Florida, USA. Association for Computational Linguistics.

Zhenyi Liao, Qingsong Xie, Yanhao Zhang, Zijian Kong, Haonan Lu, Zhenyu Yang, and Zhijie Deng. 2025. Improved visual-spatial reasoning via r1-zero-like training. *Preprint*, arXiv:2504.00883.

Fangyu Liu, Guy Emerson, and Nigel Collier. 2023. Visual spatial reasoning. *Transactions of the Association for Computational Linguistics*, 11:635–651.

Yang Liu, Ming Ma, Xiaomin Yu, Pengxiang Ding, Han Zhao, Mingyang Sun, Siteng Huang, and Donglin Wang. 2025a. SSR: Enhancing depth perception in vision-language models via rationale-guided spatial reasoning. In *Advances in Neural Information Processing Systems 38*, Advances in Neural Information Processing Systems, pages 137357–137389. Neural Information Processing Systems Foundation, Inc. (NeurIPS).

Yanli Liu, Yuting Wang, Heng Zhang, and Qi Li. 2025b. Geometric constraints and semantic optimization SLAM algorithm for dynamic scenarios. *Scientific Reports*, 15(1):31864.

Yijun Liu, Yuwei Liu, Yuan Meng, Jieheng Zhang, Yuwei Zhou, Ye Li, Jiacheng Jiang, Kangye Ji, Shijia Ge, Zhi Wang, and Wenwu Zhu. 2025c. Spatial policy: Guiding visuomotor robotic manipulation with spatial-aware modeling and reasoning. *Preprint*, arXiv:2508.15874.

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. 2025d. Spatialcot: Advancing spatial reasoning through coordinate alignment and chain-of-thought for embodied task planning. *Preprint*, arXiv:2501.10074.

Zhenyang Liu, Sixiao Zheng, Siyu Chen, Cairong Zhao, Longfei Liang, Xiangyang Xue, and Yanwei Fu. 2025e. A neural representation framework with llm-driven spatial reasoning for open-vocabulary 3d visual grounding. In *Proceedings of the 33rd ACM International Conference on Multimedia*, MM '25, page 1042–1051, New York, NY, USA. Association for Computing Machinery.

Zuntao Liu, Yi Du, Taimeng Fu, Shaoshu Su, Cherie Ho, and Chen Wang. 2026. *Vision-Language Memory for Spatial Reasoning*, page 297–316. Springer Nature Switzerland.

Zuyan Liu, Yuhao Dong, Ziwei Liu, Winston Hu, Jiwen Lu, and Yongming Rao. 2025f. Oryx MLLM: On-demand spatial-temporal understanding at arbitrary resolution. In *The Thirteenth International Conference on Learning Representations*.

Hong Lu, Hengxu Li, Prithviraj Singh Shahani, Stephanie Herbers, and Matthias Scheutz. 2025. Probing a vision-language-action model for symbolic states and integration into a cognitive architecture. In *2025 IEEE International Conference on AI and Data Analytics (ICAD)*, pages 1–8.

Wufei Ma, Haoyu Chen, Guofeng Zhang, Yu-Cheng Chou, Jieneng Chen, Celso De Melo, and Alan Yuille. 2025a. 3dsrbench: A comprehensive 3d spatial reasoning benchmark. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6924–6934.

Wufei Ma, Yu-Cheng Chou, Qihao Liu, Xingrui Wang, Celso de Melo, Jianwen Xie, and Alan Yuille. 2025b. SpatialReasoner: Towards explicit and generalizable 3D spatial reasoning. In *Advances in Neural Information Processing Systems 38*, Advances in Neural Information Processing Systems, pages 156117–156140. Neural Information Processing Systems Foundation, Inc. (NeurIPS).

Yueen Ma, Zixing Song, Yuzheng Zhuang, Jianye Hao, and Irwin King. 2026. A survey on vision–language–action models for embodied ai. *IEEE Transactions on Neural Networks and Learning Systems*, 37(7):3031–3051.

Zixian Ma, Jianguo Zhang, Zhiwei Liu, Jieyu Zhang, Juntao Tan, Manli Shu, Juan Carlos Niebles, Shelby Heinecke, Huan Wang, Caiming Xiong, Ranjay Krishna, and silvio savarese. 2025c. TACO: Learning multi-modal models to reason and act with synthetic chains-of-thought-and-action. In *Workshop on Reasoning and Planning for Large Language Models*.

Damiano Marsili, Rohun Agrawal, Yisong Yue, and Georgia Gkioxari. 2025. Visual agentic ai for spatial reasoning with a dynamic api. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19446–19455.

Julius Mayer, Mohamad Ballout, Serwan Jassim, Farbod Nosrat Nezami, and Elia Bruni. 2025. iVISPAR — an interactive visual-spatial reasoning benchmark for VLMs. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 26757–26781, Suzhou, China. Association for Computational Linguistics.

Stephanie A. McMains and David C. Somers. 2005. Processing efficiency of divided spatial attention mechanisms in human visual cortex. *Journal of Neuroscience*, 25(41):9444–9448.

Fanqing Meng, Jin Wang, Chuanhao Li, Quanfeng Lu, Hao Tian, Tianshuo Yang, Jiaqi Liao, Xizhou Zhu, Jifeng Dai, Yu Qiao, Ping Luo, Kaipeng Zhang, and Wenqi Shao. 2025. MMIU: Multimodal multi-image understanding for evaluating large vision-language models. In *The Thirteenth International Conference on Learning Representations*.

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

Roshanak Mirzaee, Hossein Rajaby Faghihi, Qiang Ning, and Parisa Kordjamshidi. 2021. SPARTQA: A textual question answering benchmark for spatial reasoning. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4582–4598, Online. Association for Computational Linguistics.

Michael J. Munje, Chen Tang, Shuijing Liu, Zichao Hu, Yifeng Zhu, Jiaxun Cui, Garrett Warnell, Joydeep Biswas, and Peter Stone. 2025. Socialnav-sub: Benchmarking vlms for scene understanding in social robot navigation. *Preprint*, arXiv:2509.08757.

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

Vienna, Austria. Association for Computational Linguistics.

Kun Ouyang. 2025. Spatial-R1: Enhancing MLLMs in video spatial reasoning. *Preprint*, arXiv:2504.01805.

Davide Paglieri, Bartłomiej Cupiał, Samuel Coward, Ulyana Piterbarg, Maciej Wolczyk, Akbir Khan, Eduardo Pignatelli, Łukasz Kuciński, Lerrel Pinto, Rob Fergus, Jakob Nicolaus Foerster, Jack Parker-Holder, and Tim Rocktäschel. 2025. Balrog: Benchmarking agentic llm and vlm reasoning on games. *Preprint*, arXiv:2411.13543.

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

Atin Pothiraj, Elias Stengel-Eskin, Jaemin Cho, and Mohit Bansal. 2025. Capture: Evaluating spatial reasoning in vision language models via occluded object counting. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8001–8010.

Tanawan Premsri and Parisa Kordjamshidi. 2025. Neuro-symbolic training for reasoning over spatial language. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 2395–2414, Albuquerque, New Mexico. Association for Computational Linguistics.

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

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. 2025b. SoFar: Language-grounded orientation bridges spatial reasoning and object manipulation. In *Advances in Neural Information Processing Systems 38*, Advances in Neural Information Processing Systems, pages 85008–85053. Neural Information Processing Systems Foundation, Inc. (NeurIPS).

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

Xingjian Ran, Yixuan Li, Linning Xu, Mulin Yu, and Bo Dai. 2025. Direct numerical layout generation for 3D indoor scene synthesis via spatial reasoning. In *Advances in Neural Information Processing Systems 38*, Advances in Neural Information Processing Systems, pages 138615–138641. Neural Information Processing Systems Foundation, Inc. (NeurIPS).

Md Imbesat Rizvi, Xiaodan Zhu, and Iryna Gurevych. 2024. SpaRC and SpaRP: Spatial reasoning characterization and path generation for understanding spatial reasoning capability of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4750–4767, Bangkok, Thailand. Association for Computational Linguistics.

Fedor Rodionov, Abdelrahman Eldesokey, Michael Birsak, John Femiani, Bernard Ghanem, and Peter Wonka. 2026. Floorplanqa: A benchmark for spatial reasoning in llms using structured representations. *Preprint*, arXiv:2507.07644.

Ayushman Sarkar, Mohd Yamani Idna Idris, and Zhenyu Yu. 2026. Reasoning in computer vision: Taxonomy, models, tasks, and methodologies. *Preprint*, arXiv:2508.10523.

Karun Sharma and Vidushee Vats. 2025. Think to ground: Improving spatial reasoning in LLMs for better visual grounding. In *Workshop on Reasoning and Planning for Large Language Models*.

Xiaoqian Shen, Yunyang Xiong, Changsheng Zhao, Lemeng Wu, Jun Chen, Chenchen Zhu, Zechun Liu, Fanyi Xiao, Balakrishnan Varadarajan, Florian Bordes, Zhuang Liu, Hu Xu, Hyunwoo J. Kim, Bilge Soran, Raghuraman Krishnamoorthi, Mohamed Elhoseiny, and Vikas Chandra. 2025a. LongVU: Spatiotemporal adaptive compression for long video-language understanding. In *Forty-second International Conference on Machine Learning*.

Yifan Shen, Yuanzhe Liu, Jingyuan Zhu, Xu Cao, Xiaofeng Zhang, Yixiao He, Wenming Ye, James Rehg, and Ismini Lourentzou. 2025b. Fine-grained preference optimization improves spatial reasoning in VLMs. In *Advances in Neural Information Processing Systems 38*, Advances in Neural Information Processing Systems, pages 20406–20437. Neural Information Processing Systems Foundation, Inc. (NeurIPS).

Zhengxiang Shi, Qiang Zhang, and Aldo Lipani. 2022. StepGame: A new benchmark for robust multi-hop spatial reasoning in texts. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11321–11329. AAAI Press.

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

Philipp D. Siedler. 2026. Sphyr: Spatial-physical reasoning benchmark on material distribution. *Preprint*, arXiv:2505.16048.

Chan Hee Song, Valts Blukis, Jonathan Tremblay, Stephen Tyree, Yu Su, and Stan Birchfield. 2025a. Robospatial: Teaching spatial understanding to 2d and 3d vision-language models for robotics. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15768–15780.

Daeun Song, Jing Liang, Amirreza Payandeh, Amir Hossain Raj, Xuesu Xiao, and Dinesh Manocha. 2025b. Vlm-social-nav: Socially aware robot navigation through scoring using vision-language models. *IEEE Robotics and Automation Letters*, 10(1):508–515.

Yonghao Song, Bingchuan Liu, Xiang Li, Nanlin Shi, Yijun Wang, and Xiaorong Gao. 2024. Decoding natural images from EEG for object recognition. In *The Twelfth International Conference on Learning Representations*.

Zijian Song, Xiaoxin Lin, Qiuming Huang, Sihan Qin, Guangrun Wang, and Liang Lin. 2026. Siri-bench: Challenging vlms' spatial intelligence through complex reasoning tasks. *Preprint*, arXiv:2506.14512.

Michael Spranger, Jakob Suchan, Mehul Bhatt, and Manfred Eppe. 2014. Grounding dynamic spatial relations for embodied (robot) interaction. In *PRICAI 2014: Trends in Artificial Intelligence*, pages 958–971, Cham. Springer International Publishing.

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

Alane Suhr, Mike Lewis, James Yeh, and Yoav Artzi. 2017. A corpus of natural language for visual reasoning. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 217–223, Vancouver, Canada. Association for Computational Linguistics.

Peiwen Sun, Shiqiang Lang, Dongming Wu, Yi Ding, Kaituo Feng, Huadai Liu, Zhen Ye, Rui Liu, Yun-Hui Liu, Jianan Wang, and Xiangyu Yue. 2026. Spacevista: All-scale visual spatial reasoning from mm to km. *Preprint*, arXiv:2510.09606.

Xuefei Sun, Doncey Albin, Cecilia Mauceri, Dusty Woods, and Christoffer Heckman. 2025. Spatial-llava: Enhancing large language models with spatial referring expressions for visual understanding. *Preprint*, arXiv:2505.12194.

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

Kexian Tang, Junyao Gao, Yanhong Zeng, Haodong Duan, Yanan Sun, Zhening Xing, Wenran Liu, Kai Chen, and Kaifeng Lyu. 2026. Lego-puzzles: How good are mllms at multi-step spatial reasoning? In *Computer Vision – ECCV 2026*, pages 1–19, Cham. Springer Nature Switzerland.

Yihong Tang, Ao Qu, Zhaokai Wang, Dingyi Zhuang, Zhaofeng Wu, Wei Ma, Shenhao Wang, Yunhan Zheng, Zhan Zhao, and Jinhua Zhao. 2025. Sparkle: Mastering basic spatial capabilities in vision language models elicits generalization to spatial reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 4083–4103, Suzhou, China. Association for Computational Linguistics.

Nazia Tasnim, Keanu Nichols, Yuting Yang, Nicholas Ikechukwu, Elva Zou, Deepti Ghadiyaram, and Bryan A. Plummer. 2026. Seeing isn't orienting: A cognitively grounded benchmark reveals systematic orientation failures in mllms supplementary. *Preprint*, arXiv:2603.11410.

Laetitia Teodorescu, Katja Hofmann, and Pierre-Yves Oudeyer. 2022. Spatialsim: Recognizing spatial configurations of objects with graph neural networks. *Frontiers in Artificial Intelligence*, 4.

Kexin Tian, Jingrui Mao, Yunlong Zhang, Jiwan Jiang, Yang Zhou, and Zhengzhong Tu. 2025. Nuscenes-spatialqa: A spatial understanding and reasoning benchmark for vision-language models in autonomous driving. In *2025 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pages 4626–4635.

Abdarahmane Traore, Éric Hervet, and Andy Couturier. 2025. Smolrgpt: Efficient spatial reasoning for warehouse environments with 600m parameters. In *2025 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pages 5330–5338.

Chengcheng Wang, Jianyuan Guo, Hongguang Li, Yuchuan Tian, Ying Nie, Chang Xu, and Kai Han. 2026a. Circle-rope: Cone-like decoupled rotary positional embedding for large vision-language models. *Preprint*, arXiv:2505.16416.

Peijie Wang, Chao Yang, Zhong-Zhi Li, Fei Yin, Dekang Ran, Mi Tian, Zhilong Ji, Jinfeng Bai, and Cheng-Lin Liu. 2025a. SolidGeo: Measuring multimodal spatial math reasoning in solid geometry. In *Advances in Neural Information Processing Systems 38*, Advances in Neural Information Processing Systems, pages 70666–70717. Neural Information Processing Systems Foundation, Inc. (NeurIPS).

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

Rong Wang and Kun Sun. 2026. Dspy-based neural-symbolic pipeline to enhance spatial reasoning in llms. *Neural Networks*, 193:108022.

Siting Wang, Minnan Pei, Luoyang Sun, Cheng Deng, Yuchen Li, Kun Shao, Zheng Tian, Haifeng Zhang, and Jun Wang. 2026b. Spatialviz-bench: A cognitively-grounded benchmark for diagnosing spatial visualization in mllms. *Preprint*, arXiv:2507.07610.

Siyu Wang, Cailian Chen, Xinyi Le, Qimin Xu, Lei Xu, Yanzhou Zhang, and Jie Yang. 2025b. CAD-GPT: Synthesising CAD construction sequence with spatial reasoning-enhanced multimodal LLMs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 7880–7888. AAAI Press.

Wenbin Wang, Yongcheng Jing, Liang Ding, Yingjie Wang, Li Shen, Yong Luo, Bo Du, and Dacheng Tao. 2025c. Retrieval-augmented perception: High-resolution image perception meets visual RAG. In *Forty-second International Conference on Machine Learning*.

Zehan Wang, Ziang Zhang, Tianyu Pang, Chao Du, Hengshuang Zhao, and Zhou Zhao. 2025d. Orient anything: Learning robust object orientation estimation from rendering 3d models. In *Forty-second International Conference on Machine Learning*.

Zhaozhi Wang, Tong Zhang, Mingyue Guo, Yaowei Wang, and Qixiang Ye. 2025e. Videoanchor: Reinforcing subspace-structured visual cues for coherent visual-spatial reasoning. *Preprint*, arXiv:2509.25151.

Azmine Toushik Wasi, Wahid Faisal, Abdur Rahman, Mahfuz Ahmed Anik, Munem Shahriar, Mohsin Mahmud Topu, Sadia Tasnim Meem, Rahatun Nesa Priti, Sabrina Afroz Mitu, Md. Iqramul Hoque, Shahriyar Zaman Ridoy, Mohammed Eunus Ali, Majd Hawasly, Mohammad Raza, and Md Rizwan Parvez. 2026a. Spatialab: Can vision-language models perform spatial reasoning in the wild? *Preprint*, arXiv:2602.03916.

Azmine Toushik Wasi, Wahid Faisal, Abdur Rahman, Mahfuz Ahmed Anik, Munem Shahriar, Mohsin Mahmud Topu, Sadia Tasnim Meem, Rahatun Nesa Priti, Sabrina Afroz Mitu, Md. Iqramul Hoque, Shahriyar Zaman Ridoy, Mohammed Eunus Ali, Majd Hawasly, Mohammad Raza, and Md Rizwan Parvez. 2026b. Spatialab: Can vision–language models perform spatial reasoning in the wild? In *The Fourteenth International Conference on Learning Representations*.

Xilin Wei, Xiaoran Liu, Yuhang Zang, Xiaoyi Dong, Pan Zhang, Yuhang Cao, Jian Tong, Haodong Duan, Qipeng Guo, Jiaqi Wang, Xipeng Qiu, and Dahua Lin. 2025. VideoroPE: What makes for good video rotary position embedding? In *Forty-second International Conference on Machine Learning*.

Changti Wu, Shijie Lian, Zihao Liu, Lei Zhang, Laurence Tianruo Yang, and Kai Chen. 2025a. Dynasolidgeo: A dynamic benchmark for genuine spatial mathematical reasoning of vlms in solid geometry. *Preprint*, arXiv:2510.22340.

Junfei Wu, Jian Guan, Kaituo Feng, Qiang Liu, Shu Wu, Liang Wang, Wei Wu, and Tieniu Tan. 2025b. Reinforcing spatial reasoning in vision-language models with interwoven thinking and visual drawing. In *Advances in Neural Information Processing Systems 38*, Advances in Neural Information Processing Systems, pages 158982–159015. Neural Information Processing Systems Foundation, Inc. (NeurIPS).

Qiong Wu, Xiangcong Yang, Yiyi Zhou, Chenxin Fang, Baiyang Song, Xiaoshuai Sun, and Rongrong Ji. 2025c. Grounded chain-of-thought for multimodal large language models. *Preprint*, arXiv:2503.12799.

Wenshan Wu, Shaoguang Mao, Yadong Zhang, Yan Xia, Li Dong, Lei Cui, and Furu Wei. 2024. Mind's eye of LLMs: Visualization-of-thought elicits spatial reasoning in large language models. In *Advances in Neural Information Processing Systems 37*, Advances in Neural Information Processing Systems, pages 90277–90317. Neural Information Processing Systems Foundation, Inc. (NeurIPS).

Jiaer Xia, Yuhang Zang, Peng Gao, Yixuan Li, and Kaiyang Zhou. 2025. Visionary-r1: Mitigating shortcuts in visual reasoning with reinforcement learning. *Preprint*, arXiv:2505.14677.

Qinghongbing Xie, Zhaoyuan Xia, Feng Zhu, Lijun Gong, Ziyue Li, Rui Zhao, and Long Zeng. 2026. Gtr-bench: Evaluating geo-temporal reasoning in vision-language models. *Preprint*, arXiv:2510.07791.

Liuchang Xu, Shuo Zhao, Qingming Lin, Luyao Chen, Qianqian Luo, Sensen Wu, Xinyue Ye, Hailin Feng, and Zhenhong Du. 2025a. Evaluating large language models on spatial tasks: A multi-task benchmarking study. *Preprint*, arXiv:2408.14438.

Xiaohao Xu, Feng Xue, Xiang Li, Haowei Li, Tianyi Zhang, Matthew Johnson-Roberson, and Xiaonan Huang. 2025b. Towards multi-hypothesis spatial foundation model: Rethinking and decoupling spatial ambiguity via laplacian visual prompting. In *ICLR 2025 Workshop on Foundation Models in the Wild*.

Qiyao Xue, Haoming Wang, Weichen Liu, Shiqi Wang, Yuyang Wu, and Wei Gao. 2026. Reasoning path and latent state analysis for multi-view visual spatial reasoning: A cognitive science perspective. In *Computer Vision – ECCV 2026*, pages 512–531, Cham. Springer Nature Switzerland.

Yuncong Yang, Jiageng Liu, Zheyuan Zhang, Siyuan Zhou, Reuben Tan, Jianwei Yang, Yilun Du, and Chuang Gan. 2025. Mindjourney: Test-time scaling with world models for spatial reasoning. In *Advances in Neural Information Processing Systems 38*, Advances in Neural Information Processing Systems, pages 121747–121777. Neural Information Processing Systems Foundation, Inc. (NeurIPS).

Wenqian Ye, Luyang Jiang, Eric Xie, Guangtao Zheng, Yunsheng Ma, Xu Cao, Dongliang Guo, Daiqing Qi, Zeyu He, Yijun Tian, Megan Coffee, Zhe Zeng, Sheng Li, Ting-hao, Huang, Ziran Wang, James M. Rehg, Henry Kautz, and Aidong Zhang. 2025. The clever hans mirage: A comprehensive survey on spurious correlations in machine learning. *Preprint*, arXiv:2402.12715.

Songsong Yu, Yuxin Chen, Hao Ju, Lianjie Jia, Fuxi Zhang, Shaofei Huang, Yuhan Wu, Rundi Cui, Binghao Ran, Zaibin Zhang, Zhedong Zheng, Zhipeng Zhang, Yifan Wang, Lin Song, Lijun Wang, Yanwei Li, Ying Shan, and Huchuan Lu. 2025. How far are vlms from visual spatial intelligence? a benchmark-driven perspective. *Preprint*, arXiv:2509.18905.

Nader Zantout, Haochen Zhang, Pujith Kachana, Jinkai Qiu, Guofei Chen, Ji Zhang, and Wenshan Wang. 2025. Sort3d: Spatial object-centric reasoning toolbox for zero-shot 3d grounding using large language models. In *2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 2201–2208.

Jirong Zha, Yuxuan Fan, Xiao Yang, Chen Gao, and Xinlei Chen. 2025. How to enable llm with 3d capacity? a survey of spatial reasoning in llm. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence, IJCAI-25*, pages 10817–10825. International Joint Conferences on Artificial Intelligence Organization. Survey Track.

Xiaoyu Zhan, Wenxuan Huang, Hao Sun, Xinyu Fu, Changfeng Ma, Shaosheng Cao, Bohan Jia, Shaohui Lin, Zhenfei Yin, Lei Bai, Wanli Ouyang, Yuanqi Li, Jie Guo, and Yanwen Guo. 2025. Actial: Activate spatial reasoning ability of multimodal large language models. In *Advances in Neural Information Processing Systems 38*, Advances in Neural Information Processing Systems, pages 93542–93567. Neural Information Processing Systems Foundation, Inc. (NeurIPS).

Ji Zhang, Shihan Wu, Xu Luo, Hao Wu, Lianli Gao, Heng Tao Shen, and Jingkuan Song. 2025a. Inspire: Vision-language-action models with intrinsic spatial reasoning. *Preprint*, arXiv:2505.13888.

Jianke Zhang, Yanjiang Guo, Yucheng Hu, Xiaoyu Chen, Xiang Zhu, and Jianyu Chen. 2025b. UP-VLA: A unified understanding and prediction model for embodied agent. In *Forty-second International Conference on Machine Learning*.

Kevin Zhang, Kuangzhi Ge, Xiaowei Chi, Renrui Zhang, Shaojun Shi, Zhen Dong, Sirui Han, and Shanghang Zhang. 2025c. Can world models benefit vlms for world dynamics? *Preprint*, arXiv:2510.00855.

Qianheng Zhang, Song Gao, Chen Wei, Yibo Zhao, Ying Nie, Ziru Chen, Shijie Chen, Yu Su, and Huan Sun. 2025d. Geoanalystbench: A geoai benchmark for assessing large language models for spatial analysis workflow and code generation. *Transactions in GIS*, 29(7):e70135. E70135 8156745.

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

Weichen Zhang, Zile Zhou, Xin Zeng, Liu Xuchen, Jianjie Fang, Chen Gao, Jinqiang Cui, Yong Li, Xinlei Chen, and Xiao-Ping Zhang. 2025f. Open3d-vqa: A benchmark for embodied spatial concept reasoning with multimodal large language model in open space. In *Proceedings of the 33rd ACM International Conference on Multimedia*, MM '25, page 12784–12791, New York, NY, USA. Association for Computing Machinery.

Wenyu Zhang, Wei En Ng, Lixin Ma, Yuwen Wang, Jungqi Zhao, Boyang Li, and Lu Wang. 2024. SPHERE: A hierarchical evaluation on spatial perception and reasoning for vision-language models. *Preprint*, arXiv:2412.12693. Version v1, 17 Dec 2024; License: CC BY 4.0.

Yi Zhang, Qiang Zhang, Xiaozhu Ju, Zhaoyang Liu, Jilei Mao, Jingkai Sun, Jintao Wu, Shixiong Gao, Shihan Cai, Zhiyuan Qin, Linkai Liang, Jiaxu Wang, Yiqun Duan, Jiahang Cao, Renjing Xu, and Jian Tang. 2025g. Embodiedvsr: Dynamic scene graph-guided chain-of-thought reasoning for visual spatial tasks. *Preprint*, arXiv:2503.11089.

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

Yuyou Zhang, Radu Corcodel, Chiori Hori, Anoop Cherian, and Ding Zhao. 2026. Spinbench: Perspective and rotation as a lens on spatial reasoning in vlms. *Preprint*, arXiv:2509.25390.

Zefeng Zhang, Xiangzhao Hao, Hengzhu Tang, Zhenyu Zhang, Jiawei Sheng, Xiaodong Li, Zhenyang Li, Li Gao, Daiting Shi, Dawei Yin, and Tingwen Liu. 2025i. Cooper: A unified model for cooperative perception and reasoning in spatial intelligence. *Preprint*, arXiv:2512.04563.

Ziang Zhang, Zehan Wang, Guanghao Zhang, Weilong Dai, Yan Xia, Ziang Yan, Minjie Hong, and Zhou Zhao. 2025j. Dsi-bench: A benchmark for dynamic spatial intelligence. *Preprint*, arXiv:2510.18873.

Baining Zhao, Ziyou Wang, Jianjie Fang, Chen Gao, Fanhang Man, Jinqiang Cui, Xin Wang, Xinlei Chen, Yong Li, and Wenwu Zhu. 2025. Embodied-r: Collaborative framework for activating embodied spatial reasoning in foundation models via reinforcement learning. In *Proceedings of the 33rd ACM International Conference on Multimedia*, MM '25, page 11071–11080, New York, NY, USA. Association for Computing Machinery.

Pukun Zhao, Longxiang Wang, Miaowei Wang, Chen Chen, Fanqing Zhou, and Haojian Huang. 2026a. EvoEmpirBench: Dynamic spatial reasoning with Agent-ExpVer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 36564–36572. AAAI Press.

Yu Zhao, Wei-Ning Chen, Huseyin Atahan Inan, Samuel Kessler, Lu Wang, Lukas Wutschitz, Fangkai Yang, Chaoyun Zhang, Pasquale Minervini, Saravan Rajmohan, and Robert Sim. 2026b. Learning gui grounding with spatial reasoning from visual feedback. *Preprint*, arXiv:2509.21552.

Enshen Zhou, Jingkun An, Cheng Chi, Yi Han, Shanyu Rong, Chi Zhang, Pengwei Wang, Zhongyuan Wang, Tiejun Huang, Lu Sheng, and Shanghang Zhang. 2025a. RoboRefer: Towards spatial referring with

reasoning in vision-language models for robotics. In *Advances in Neural Information Processing Systems 38*, Advances in Neural Information Processing Systems, pages 32209–32286. Neural Information Processing Systems Foundation, Inc. (NeurIPS).

Shijie Zhou, Alexander Vilesov, Xuehai He, Ziyu Wan, Shuwang Zhang, Aditya Nagachandra, Di Chang, Dongdong Chen, Xin Eric Wang, and Achuta Kadambi. 2025b. Vlm4d: Towards spatiotemporal awareness in vision language models. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8600–8612.

Xingyi Zhou, Anurag Arnab, Chen Sun, and Cordelia Schmid. 2025c. Dense video object captioning from disjoint supervision. In *The Thirteenth International Conference on Learning Representations*.

A Related Work

A growing body of survey and position papers has examined spatial reasoning in language, vision–language, and embodied models, but existing treatments are often fragmented along modality or task boundaries. Early analyses focused on diagnostic datasets for relational and compositional reasoning in vision–language models, highlighting shortcut learning and the gap between recognition and relational understanding (Song et al., 2024; Du et al., 2024). More recent surveys and benchmark-driven studies emphasize empirical failures of large models on spatial tasks despite strong headline performance, particularly for metric reasoning, perspective changes, and multi-hop relational inference (Ramakrishnan et al., 2025; Zhao et al., 2026a; Huang et al., 2025b). While these works provide valuable empirical insights, they typically organize the literature around benchmarks or model classes rather than offering a unified conceptual framework that connects tasks, relations, reasoning processes, and internal representations. As a result, comparisons across modalities and evaluations often conflate task difficulty with genuine spatial competence.

Our work differs from prior surveys by proposing a unified, multi-axis taxonomy of spatial reasoning explicitly designed to bridge NLP, computer vision, and embodied AI. Instead of cataloging benchmarks or architectures in isolation, we decompose spatial reasoning along complementary dimensions, task formulation, relation type, reasoning process, representation, and temporal–memory demands, allowing systematic comparison of models and evaluations that would otherwise appear incomparable. This perspective builds on empirical critiques of shortcut learning and weak grounding (Ramakrishnan et al., 2025; Huang et al., 2025b), but goes further by explaining *why* particular failure modes arise and *which* aspects of spatial reasoning they reflect. In contrast to benchmark-centric reviews (Zhao et al., 2026a), we explicitly disentangle symbolic consistency from geometric grounding, static perception from dynamic and agentic reasoning, and implicit latent encoding from structured or world-model-based representations. By synthesizing results across text-only, visual, 3D, and embodied settings within a single taxonomy, our survey exposes blind spots in current evaluation practice and clarifies what progress toward genuine spatial intelligence should entail. We believe this integrative framing is essential for guiding future benchmark design, architectural choices, and training strategies beyond incremental gains on existing datasets.

B Details on Benchmarks and Datasets

Spatial reasoning benchmarks have evolved from early, controlled diagnostics to more complex tests involving 3D structure, temporal continuity, and agentic interaction, yet progress remains uneven due to shortcut learning and limited relational scope (Johnson et al., 2017; Hudson and Manning, 2019; Yu et al., 2025; Pothiraj et al., 2025; Wang and Ling, 2025). Here, we categorize major benchmarks into three phases to assess which spatial reasoning aspects they truly evaluate and where they systematically fail.

★ **Phase I: Static and Diagnostic Benchmarks.** Phase I benchmarks isolate spatial and relational reasoning in controlled settings, often with synthetic scenes or constrained language. Datasets like CLEVR and NLVR2 focus on compositional question answering about object attributes and relations (Johnson et al., 2017; Hudson and Manning, 2019; Yu et al., 2025; Deng et al., 2025), while text-centric benchmarks such as SpartQA and StepGame emphasize multi-hop relational consistency and symbolic inference (Stogiannidis et al., 2025; Tasnim et al., 2026). Despite early relational reasoning gains, many models exploited dataset biases rather than forming explicit spatial representations, making Phase I performance a weak predictor of real-world generalization (Suhr et al., 2017; Wang and Ling, 2025; Pothiraj et al., 2025; Ao et al., 2025).

★ **Phase II: Cognitive and 3D-Centric Benchmarks.** Phase II benchmarks address static 2D limitations by incorporating explicit 3D structure, viewpoint changes, and cognitive spatial tasks. Datasets like Open3DVQA and SURPRISE3D test volumetric understanding, occlusion, and cross-view consistency (Pothiraj et al., 2025; Danish et al., 2025), while SPARE3D and OmniSpatial focus on mental rotation and metric consistency inspired by human spatial cognition (Stogiannidis et al., 2025; Munje et al., 2025). Models often show sharp declines on Phase II tasks compared to Phase I, particularly on metric and perspective-dependent queries, highlighting the necessity of 3D-centric evaluation (Chen et al., 2024; Wang and Ling, 2025).

★ **Phase III: Dynamic and Agentic Benchmarks.**

Table 1: Spatial reasoning types in LLMs and VLMs with example questions illustrating each subtype.

Type	Sub-Type	Example
Task-Level Spatial Reasoning	Text-Only Spatial Reasoning	"A cup is on the table. There is a flower under the table. Where is the flower relative to the cup?"
	Image-Based Spatial Reasoning	"In this image, a red cube is to the left of a blue sphere, which is behind a green pyramid. What object is closest to the viewer?" (CLEVR)
	Video and Spatiotemporal Reasoning	"A car moves from the left side of the frame to the right, passing behind a tree. Where is the car now relative to the tree?" (VLM4D)
	3D and Multi-View Spatial Reasoning	"Given two images of a chair from different angles, is the armrest on the right side consistent across views?" (Open3DVQA)
	Embodied and Agentic Spatial Reasoning	"You are in a room facing north. The door is behind you. You turn right and walk forward. Where is the door now relative to you?"
Relation-Level Spatial Reasoning	Topological Relations	"Is the book inside the box or on top of it?" (GQA)
	Directional Relations	"If the tree is to the east of the house, and the bench is south of the tree, where is the bench relative to the house?" (Winoviz)
	Metric and Quantitative Relations	"Is the red ball larger than the blue ball?" (MapEval)
	Perspective-Dependent Relations	"If you rotate a shape 90 degrees clockwise, which side will now face north?" (iVISPAR's Sliding Geom Puzzle)
	Relational Composition	"A is left of B, B is in front of C. Where is A relative to C?" (SPaRC)
Reasoning Processes	Single-Hop Reasoning	"Is the cup on the table?"
	Multi-Hop and Sequential Reasoning	"The book is on the shelf. The shelf is above the table. Where is the book relative to the table?"
	Compositional Reasoning	"The blue cube is left of the red sphere and behind the green pyramid. What is the position of the blue cube relative to the green pyramid?"
	Dynamic and Predictive Reasoning	"If the ball rolls off the table, where will it land?" (FoundationMotion)
	Counterfactual and Hypothetical Reasoning	"If the cup were placed under the table instead of on it, where would it be relative to the flower?"

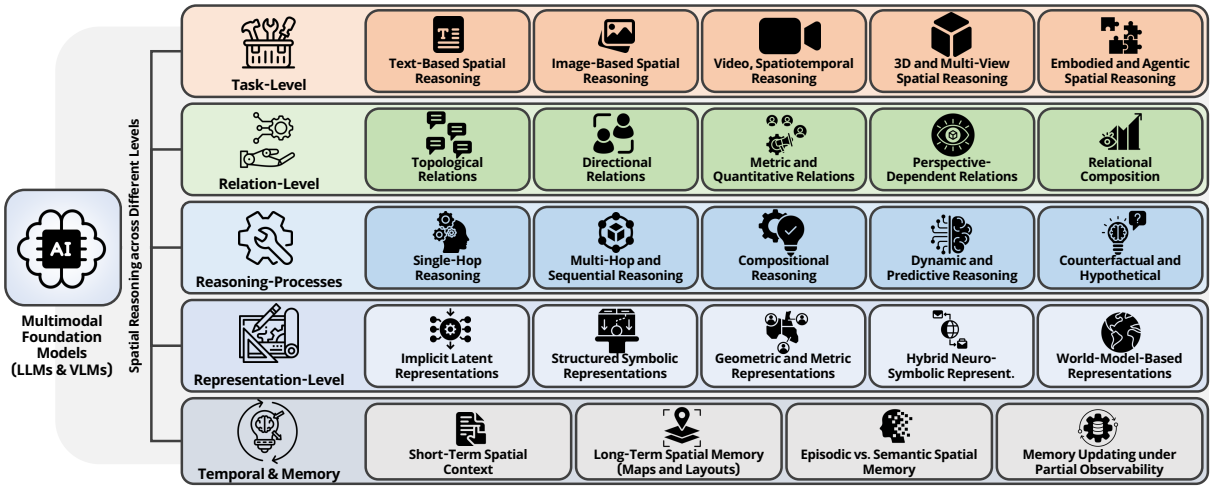


Figure 2: Spatial reasoning across multiple levels, from tasks to underlying processes and dynamics.

Table 2: Glossary of spatial reasoning terms and technical jargon for general understanding.

Term	Explanation
Text-Only Spatial Reasoning	Reasoning about spatial relationships using only written or spoken language without visual input.
Image-Based Spatial Reasoning	Reasoning about spatial relationships based on static images.
Video and Spatiotemporal Reasoning	Understanding spatial relationships that change over time in videos or sequences.
3D and Multi-View Spatial Reasoning	Spatial reasoning involving depth, volume, and different viewpoints, often using multiple images or 3D data.
Embodied and Agentic Spatial Reasoning	Spatial reasoning performed by an agent that perceives and acts in a physical or simulated environment.
Topological Relations	Qualitative relations like containment, adjacency, and connectivity that don't depend on exact measurements.
Directional Relations	Spatial relations defined by direction (e.g., left, right, north), often viewpoint-dependent.
Metric and Quantitative Relations	Precise measurements such as distances, angles, and sizes between objects.
Perspective-Dependent Relations	Relations that change when viewed from different angles, requiring mental transformations like rotation.
Relational Composition	Combining multiple spatial relations to understand complex configurations or infer new relations.
Single-Hop Reasoning	Simple spatial reasoning requiring one direct inference step.
Multi-Hop and Sequential Reasoning	Complex reasoning involving multiple inference steps chained together.
Compositional Reasoning	Reasoning that integrates various spatial concepts and object attributes simultaneously.
Dynamic and Predictive Reasoning	Forecasting how spatial relations evolve over time or predicting future states.
Counterfactual and Hypothetical Reasoning	Considering "what-if" scenarios involving changes in spatial arrangements and reasoning about their consequences.
Viewpoint Invariance	The ability to recognize spatial relations regardless of the observer's angle or perspective.
Occlusion	When one object blocks part or all of another object from view, complicating spatial reasoning.
Scene Graph	A structured representation that models objects as nodes and spatial or semantic relationships as edges.
Chain-of-Thought (CoT) Prompting	A method to encourage models to perform multi-step reasoning by explicitly generating intermediate steps.
Slot-Based Memory	A technique to store and manipulate discrete objects or entities separately in a model's internal state.
Spatial Shortcuts	Simplified or biased cues models might rely on instead of genuine spatial understanding, often exploiting dataset biases.
Egocentric vs. Allocentric Frames	Egocentric: spatial relations relative to the observer (e.g., left/right of me). Allocentric: relations relative to external references (e.g., north/south).
Metric Loss	A training loss function that encourages accurate prediction of spatial measurements like distance or depth.
Viewpoint Alignment	Adjusting or normalizing data so that different views or perspectives can be compared consistently.
Symbolic Planner	A system that uses explicit symbolic representations and rules to plan actions or reason spatially.
World Model	An internal representation that simulates the environment's dynamics and spatial properties for prediction or planning.
Replanning	The ability to revise a plan when new spatial information or errors occur.
Episodic Memory	Memory system that stores specific past experiences or states for recall during reasoning or planning.
Reinforcement Learning (RL)	A method where agents learn to act by receiving rewards or penalties, useful for embodied spatial tasks.
Multi-View Fusion	Combining spatial information from multiple viewpoints into a coherent understanding.
Mental Rotation	The cognitive ability to imagine rotating objects mentally to understand their spatial configuration.
Contrastive Learning	A training technique that teaches models to distinguish between similar and dissimilar spatial arrangements.
Spatial Prompting	Using spatially informative prompts to guide models towards better spatial reasoning.
Compositional Generalization	The ability to apply learned spatial concepts in new combinations not seen during training.

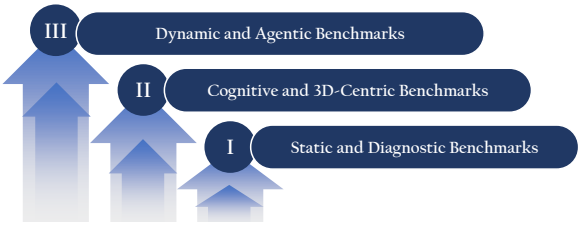


Figure 3: Three phases for benchmarks and datasets on spatial reasoning.

Phase III benchmarks situate models in dynamic, interactive environments requiring integration of spatial reasoning with perception, memory, and action. Agentic tasks in iVISPAR and BALROG assess navigation, manipulation, and long-term spatial state updates (Song et al., 2025a; Zhang et al., 2025e), while video benchmarks like VLM4D test reasoning about motion and temporal causality (Danish et al., 2025; Zhang et al., 2025e). These benchmarks reveal error compounding where small failures escalate, and success on earlier phases rarely translates here, emphasizing the critical need for dynamic evaluations (Pothiraj et al., 2025; Song et al., 2025a; Munje et al., 2025).

Mapping these phases onto the framework in §3 shows Phase I emphasizes topological and shallow compositional reasoning but neglects metric and temporal aspects (Hudson and Manning, 2019; Wang and Ling, 2025). Phase II expands to perspective and metric relations but often misses long-horizon memory demands (Pothiraj et al., 2025; Danish et al., 2025). Phase III addresses these gaps but adds complexity in control and language grounding (Song et al., 2025a; Zhang et al., 2025e). Across all, counterfactual spatial reasoning remains underexplored despite its importance for planning and imagination (Stogiannidis et al., 2025; Tasnim et al., 2026), underscoring the need for coordinated, framework-aware benchmark suites.

C Evaluation Protocols and Metrics

Evaluation protocols play a decisive role in shaping conclusions about spatial reasoning capabilities in foundation models. While benchmark accuracy remains the dominant metric, growing evidence suggests that it often masks fundamental weaknesses in representation, generalization, and reasoning process (Hudson and Manning, 2019; Pothiraj et al., 2025). Recent work has therefore advocated for diagnostic, counterfactual, and interactive evaluation paradigms that more closely align with the cognitive demands of spatial reasoning (Stogiannidis et al., 2025; Tasnim et al., 2026; Song et al., 2025a). In this section, we critically examine existing evaluation practices and discuss their implications for interpreting reported performance.

★ **Accuracy-Based Metrics and Their Limits.** Accuracy-based metrics, such as overall question-answering correctness, are widely used due to their simplicity and comparability across models. However, these metrics conflate surface-level pattern matching with genuine spatial understanding, particularly in datasets where shortcuts and biases are prevalent (Johnson et al., 2017; Wang and Ling, 2025). High accuracy can often be achieved without maintaining spatial consistency across multiple queries or viewpoints, obscuring underlying failure modes (Pothiraj et al., 2025; Suhr et al., 2017). Moreover, accuracy does not distinguish between errors arising from perceptual noise and those caused by flawed spatial reasoning (Hudson and Manning, 2019; Danish et al., 2025). As a result, sole reliance on accuracy metrics risks overstating progress in spatial reasoning, especially when models are evaluated in isolation rather than under systematic perturbation (Pothiraj et al., 2025).

★ **Diagnostic and Counterfactual Evaluation.** Diagnostic evaluations aim to isolate specific reasoning skills by controlling for confounding factors such as language priors or visual saliency (Izadi et al., 2025). Counterfactual evaluation, in particular, tests whether a model's predictions change appropriately under minimal but semantically meaningful spatial modifications (Stogiannidis et al., 2025; Tasnim et al., 2026). Studies applying counterfactual scene edits or relational swaps have revealed that many models fail to update their answers consistently, indicating brittle internal representations (Pothiraj et al., 2025; Wang and Ling, 2025). Such protocols are especially effective at exposing shortcut learning in image-based benchmarks (Johnson et al., 2017; Suhr et al., 2017). Despite their diagnostic value, counterfactual evaluations remain underutilized due to their higher annotation and implementation cost (Danish et al., 2025; Hudson and Manning, 2019). Expanding their use is essential for disentangling true spatial reasoning from dataset exploitation.

★ **Agentic and Interactive Metrics.** Agentic evaluation metrics assess spatial reasoning in the context of sequential decision-making, where errors compound over time. Metrics such as navigation success rate, path efficiency, and task completion

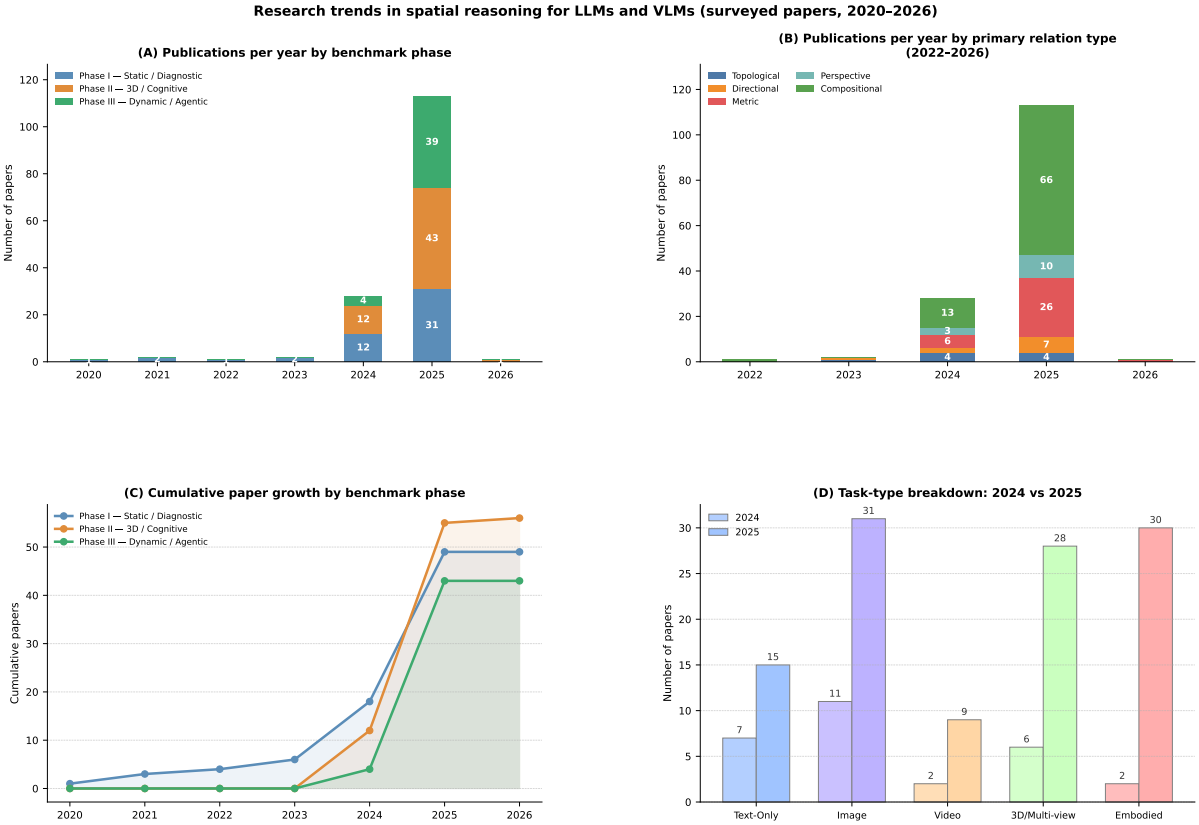


Figure 4: Publication volume and benchmark type distribution for spatial reasoning in LLMs/VLMs (2022–2025), organized by the taxonomy dimensions in §3. The trend shows a shift from Phase I (static/diagnostic) toward Phase II and Phase III (3D-centric and agentic) benchmarks from 2024 onward, with metric reasoning and embodied settings gaining the most coverage.

under partial observability provide a more stringent test of spatial competence than static accuracy alone (Song et al., 2025a; Zhang et al., 2025e). Interactive benchmarks further enable evaluation of memory consistency, replanning ability, and robustness to environmental changes (Pothiraj et al., 2025; Song et al., 2025a). Empirical results consistently show that models with strong static benchmark performance often degrade rapidly in interactive settings (Munje et al., 2025; Zhang et al., 2025e). These findings suggest that agentic metrics capture dimensions of spatial reasoning that are invisible to traditional evaluation protocols (Pothiraj et al., 2025; Song et al., 2025a).

★ **Human Baselines and Cognitive Validity.** Human baselines provide a critical reference point for interpreting spatial reasoning performance, particularly on cognitively demanding tasks such as mental rotation and perspective-taking. Comparisons with human performance reveal not only quantitative gaps but also qualitative differences in error patterns and generalization behavior (Stogiannidis et al., 2025; Tasnim et al., 2026). In many benchmarks, models fail on instances that humans find trivial, while succeeding on others that rely on statistical regularities rather than reasoning (Pothiraj et al., 2025; Munje et al., 2025). Incorporating cognitively grounded task design and human-inspired evaluation criteria can therefore improve the ecological validity of spatial reasoning benchmarks (Stogiannidis et al., 2025; Danish et al., 2025; Lu et al., 2025). Without such baselines, it remains difficult to assess whether observed improvements reflect genuine progress toward spatial intelligence or merely improved exploitation of benchmark structure (Hudson and Manning, 2019; Wang and Ling, 2025).

D Architectural Approaches to Spatial Reasoning

Architectural design plays a central role in determining whether spatial reasoning emerges as a robust capability or remains an artifact of dataset bias. While recent foundation models exhibit impressive multimodal performance, their spatial competence is tightly coupled to representational choices made

at the perception, encoding, and reasoning stages (Pothiraj et al., 2025; Chen et al., 2024). In this section, we analyze major architectural strategies for spatial reasoning, emphasizing how inductive biases, encoding schemes, and internal simulation mechanisms influence generalization and failure modes. Rather than surveying models exhaustively, we focus on architectural patterns that recur across systems and benchmark evaluations.

★ **Vision Encoders and Spatial Biases.** Most vision-language models rely on generic 2D vision encoders, such as convolutional neural networks or Vision Transformers, originally designed for object recognition rather than spatial inference. These encoders emphasize local texture and global semantic aggregation, but provide limited inductive bias for metric structure, depth, or viewpoint invariance (Johnson et al., 2017; Pothiraj et al., 2025). As a result, spatial relations are often inferred implicitly from appearance correlations rather than explicit geometric reasoning (Wang and Ling, 2025; Suhr et al., 2017). Empirical studies show that replacing CNNs with ViTs improves relational flexibility but does not resolve failures in distance estimation or 3D alignment (Hudson and Manning, 2019; Danish et al., 2025). Furthermore, pretraining on large-scale 2D image–text corpora reinforces planar biases that degrade performance on volumetric or embodied benchmarks (Munje et al., 2025; Song et al., 2025a). These limitations suggest that vision encoders optimized for recognition alone are insufficient foundations for robust spatial reasoning.

★ **Positional and Geometric Encoding.** Positional encoding schemes aim to inject spatial structure into otherwise permutation-invariant architectures, particularly Transformers. Standard absolute or relative positional encodings capture token order but provide only weak geometric grounding when applied to visual patches (Chen et al., 2024; Pothiraj et al., 2025). Extensions such as rotational or circular positional encodings attempt to improve orientation awareness, yet remain limited to 2D transformations and struggle with depth or scale variation (Danish et al., 2025; Munje et al., 2025). Depth-aware and geometry-informed encodings have been shown to improve performance on 3D-centric benchmarks, but are rarely integrated into end-to-end multimodal models (Chen et al., 2024; Wang and Ling, 2025). Moreover, positional encodings alone do not enforce global spatial consistency, leading to contradictions across queries or viewpoints (Pothiraj et al., 2025; Hudson and Manning, 2019). This suggests that positional information must be coupled with higher-level structural constraints to support reliable spatial reasoning.

★ **Scene Graphs and Structured Perception.** Structured perception approaches explicitly represent objects and their relations using intermediate symbolic or graph-based representations (Teodorescu et al., 2022). Scene graphs, in particular, provide a compositional abstraction that aligns naturally with relational spatial queries (Hudson and Manning, 2019; Tasnim et al., 2026). Empirical evidence indicates that models leveraging explicit object–relation structure exhibit improved consistency on compositional and multi-hop spatial reasoning tasks (Suhr et al., 2017; Wang and Ling, 2025). However, scene graph construction often relies on external detectors and heuristics, introducing brittleness and limiting scalability (Pothiraj et al., 2025). Additionally, fixed relational schemas may fail to capture continuous or metric spatial properties critical for navigation and manipulation (Chen et al., 2024; Song et al., 2025a). Consequently, while structured perception enhances interpretability and compositionality, it remains an incomplete solution for end-to-end spatial reasoning.

★ **World Models and 3D Internal Simulation.** World-model-based architectures aim to endow agents with internal representations that support prediction, planning, and counterfactual reasoning over spatial dynamics. Video prediction models and latent dynamical systems learn implicit simulators of environment evolution, enabling limited foresight over future spatial states (Zhang et al., 2025e; Danish et al., 2025). Recent work integrates such models with language and vision, allowing agents to reason about trajectories and interactions beyond immediate observations (Song et al., 2025a; Munje et al., 2025). Despite their promise, these models remain computationally expensive and often struggle with long-horizon consistency and error accumulation (Pothiraj et al., 2025; Zhang et al., 2025e). Moreover, learned world models frequently encode appearance dynamics more faithfully than geometric structure, limiting their utility for precise spatial planning (Chen et al., 2024; Wang and Ling, 2025). Bridging the gap between predictive simulation and explicit geometric reasoning remains a key open challenge for spatially grounded AI systems.

E Training Paradigms for Spatial Reasoning

Training strategy is a decisive factor in whether spatial reasoning emerges as a robust capability or remains a fragile byproduct of scale. While architectural inductive biases shape representational capacity, training data, supervision signals, and learning curricula determine which spatial abstractions are actually acquired (Chen et al., 2024; Pothiraj et al., 2025). Empirical evidence increasingly suggests that naive scaling of data and parameters is insufficient for learning metric, perspective-dependent, or long-horizon spatial reasoning (Wang and Ling, 2025; Danish et al., 2025). This section reviews major training paradigms proposed to address these limitations and analyzes their effectiveness across benchmark categories.

★ **Synthetic Data and Programmatic Generation.** Synthetic datasets and programmatically generated environments have played a central role in early progress on spatial reasoning, offering precise control over relational structure and supervision (Johnson et al., 2017; Hudson and Manning, 2019; Ogezi and Shi, 2025). By eliminating natural image biases, such data enable targeted evaluation of compositional and multi-hop spatial reasoning (Tasnim et al., 2026; Stogiannidis et al., 2025). However, models trained extensively on synthetic data often exhibit poor transfer to real-world scenes, suggesting overfitting to simplified visual and relational distributions (Pothiraj et al., 2025; Wang and Ling, 2025). Moreover, synthetic environments typically underspecify sensor noise, occlusion, and metric ambiguity, limiting their ability to induce robust geometric understanding (Danish et al., 2025; Chen et al., 2024). While synthetic data remains valuable for diagnostic training and evaluation, it is insufficient as a standalone strategy for real-world spatial reasoning (Pothiraj et al., 2025; Hudson and Manning, 2019).

★ **3D Lifting and Metric Supervision.** 3D lifting approaches aim to bridge the gap between 2D perception and volumetric reasoning by supervising models with depth, point clouds, or explicit 3D structure. Incorporating metric supervision has been shown to improve performance on distance estimation, viewpoint invariance, and occlusion-aware reasoning tasks (Chen et al., 2024; Danish et al., 2025). However, such supervision is expensive to obtain at scale and often limited to narrow domains such as indoor scenes or simulated environments (Pothiraj et al., 2025; Munje et al., 2025). Additionally, many models fail to propagate metric consistency into downstream reasoning modules, treating 3D cues as auxiliary features rather than core representational constraints (Wang and Ling, 2025; Hudson and Manning, 2019). These limitations suggest that metric supervision must be tightly integrated with reasoning objectives to yield durable spatial competence (Chen et al., 2024; Pothiraj et al., 2025).

★ **Instruction Tuning and Curriculum Learning.** Instruction tuning has emerged as a powerful mechanism for eliciting latent reasoning abilities in large models, including spatial inference (Jia et al., 2026a; Hudson and Manning, 2019). Carefully designed curricula that progress from simple topological relations to complex compositional and perspective-dependent tasks can significantly improve spatial consistency (Stogiannidis et al., 2025; Tasnim et al., 2026). Nevertheless, improvements from instruction tuning are often brittle, with performance degrading sharply outside the tuned distribution (Pothiraj et al., 2025; Wang and Ling, 2025). Curriculum effects also depend strongly on task ordering and representation alignment, making them difficult to generalize across benchmarks (Danish et al., 2025; Hudson and Manning, 2019). As a result, instruction tuning appears more effective at surfacing existing capabilities than at instilling fundamentally new spatial abstractions (Chen et al., 2024; Pothiraj et al., 2025).

★ **Reinforcement Learning and Tool Use.** RL introduces action-conditioned feedback that can, in principle, ground spatial reasoning in interaction and consequence (Wu et al., 2025b). In embodied settings, RL has been shown to improve navigation, spatial memory, and affordance learning relative to purely supervised training (Song et al., 2025a; Zhang et al., 2025e; Ji et al., 2025; Xia et al., 2025). However, RL-trained agents often struggle with sample efficiency and generalization, particularly when environments differ from training conditions (Pothiraj et al., 2025; Danish et al., 2025). Tool-augmented approaches, such as external planners or spatial simulators, can partially mitigate these issues by offloading precise computation and state tracking (Chen et al., 2024; Song et al., 2025a). Yet reliance on tools raises questions about where spatial reasoning resides, within the model or in the surrounding system (Wang and Ling, 2025; Hudson and Manning, 2019). Balancing internal spatial competence with external tool use remains an open

challenge for scalable agentic learning.

F Reasoning-Time Augmentation

Beyond architecture and training, a growing line of work explores *reasoning-time* interventions that modify how models perform inference without changing their underlying parameters. These approaches aim to elicit more structured spatial reasoning by encouraging explicit decomposition, intermediate representations, or external computation (Jia et al., 2026a; Hudson and Manning, 2019). Empirical evidence suggests that such interventions can significantly improve performance on diagnostic spatial tasks, although gains are often uneven and task-dependent (Pothiraj et al., 2025; Wang and Ling, 2025). This section reviews three major categories of reasoning-time augmentation and analyzes their strengths and limitations in the context of spatial reasoning.

★ **Chain-of-Thought and Spatial Decomposition.** Chain-of-thought prompting encourages models to externalize intermediate reasoning steps, which has been shown to improve multi-hop and compositional spatial reasoning in both text-only and multimodal settings (Jia et al., 2026a; Stogiannidis et al., 2025; Wu et al., 2025c; Ji et al., 2025). By decomposing complex spatial queries into simpler relational subproblems, models are better able to maintain consistency across intermediate states (Tasnim et al., 2026; Hudson and Manning, 2019). However, improvements from chain-of-thought are often superficial, as models may generate plausible but incorrect reasoning traces that are not grounded in internal representations (Pothiraj et al., 2025; Wang and Ling, 2025). Moreover, the effectiveness of spatial decomposition is highly sensitive to prompt design and task structure, limiting reproducibility across benchmarks (Danish et al., 2025; Hudson and Manning, 2019). These findings suggest that while chain-of-thought can reveal latent reasoning capacity, it does not reliably induce new spatial abstractions (Chen et al., 2024; Pothiraj et al., 2025).

★ **Visualization-of-Thought and Layout Reasoning.** Visualization-based reasoning-time interventions extend chain-of-thought by encouraging models to construct intermediate spatial layouts, diagrams, or symbolic representations. Approaches that prompt models to describe or manipulate abstract layouts have demonstrated improved performance on diagrammatic and relational spatial tasks (Hudson and Manning, 2019; Tasnim et al., 2026). Such visualizations can act as external memory, reducing the burden on implicit latent representations and improving compositional consistency (Suhr et al., 2017; Wang and Ling, 2025). Nonetheless, these methods often rely on textual proxies for geometry, limiting their effectiveness for metric or perspective-dependent reasoning (Chen et al., 2024; Pothiraj et al., 2025). Additionally, generated layouts are rarely validated for geometric correctness, allowing subtle inconsistencies to propagate through the reasoning process (Danish et al., 2025; Munje et al., 2025). As a result, visualization-of-thought remains a promising but fragile augmentation strategy.

★ **Tool-Augmented Spatial Reasoning.** Tool-augmented reasoning integrates external modules such as spatial solvers, simulators, or symbolic planners into the inference loop. By delegating precise geometric computation to specialized tools, models can overcome limitations in metric estimation and long-horizon consistency (Chen et al., 2024; Song et al., 2025a). Empirical studies show substantial gains on navigation, manipulation, and multi-step spatial planning tasks when tool use is permitted (Zhang et al., 2025e; Song et al., 2025a). However, reliance on tools complicates evaluation, as performance reflects system-level competence rather than intrinsic spatial reasoning ability (Pothiraj et al., 2025; Wang and Ling, 2025). Furthermore, models often struggle to decide when and how to invoke tools appropriately, leading to brittle behavior under distribution shift (Danish et al., 2025; Hudson and Manning, 2019). These trade-offs raise fundamental questions about whether tool-augmented success represents progress toward spatial intelligence or a pragmatic workaround for architectural limitations (Chen et al., 2024; Pothiraj et al., 2025).

G Spatial Reasoning for Embodied and Agentic AI

Embodied and agentic settings represent the most stringent test of spatial reasoning, as they require models to integrate perception, memory, reasoning, and action over extended time horizons. Unlike static benchmarks, embodied tasks expose compounding errors and demand continual state updating under partial observability (Song et al., 2025a; Zhang et al., 2025e). Empirical evidence increasingly shows that success on image-based or 3D-

centric benchmarks does not reliably translate to effective embodied behavior (Pothiraj et al., 2025; Munje et al., 2025). This section examines how spatial reasoning manifests in navigation, manipulation, and interactive world modeling, highlighting both progress and persistent limitations.

★ **Navigation and Manipulation.** Navigation and manipulation tasks require agents to reason about spatial layouts, object affordances, and action consequences in a closed perception–action loop. Benchmarks in this domain reveal that many agents struggle to maintain consistent spatial maps, leading to disorientation and inefficient paths (Song et al., 2025a; Zhang et al., 2025e). Manipulation further compounds these challenges by introducing fine-grained metric constraints and contact dynamics that are poorly captured by implicit visual representations (Chen et al., 2024; Pothiraj et al., 2025). Even when language instructions are understood correctly, agents often fail to ground them in executable spatial plans (Munje et al., 2025; Hudson and Manning, 2019). These failures suggest that embodied spatial reasoning requires stronger integration between geometric representation and policy learning (Danish et al., 2025; Song et al., 2025a).

★ **Vision–Language–Action Models.** Vision–language–action (VLA) models aim to unify perception, reasoning, and control within a single framework, leveraging language as an intermediate planning and abstraction tool (Song et al., 2025a; Zhang et al., 2025e). While such models demonstrate improved flexibility and instruction-following, their spatial reasoning remains fragile, particularly under novel configurations or environmental perturbations (Pothiraj et al., 2025; Munje et al., 2025). Language supervision often emphasizes semantic correctness over geometric feasibility, leading to plans that are linguistically coherent but spatially invalid (Hudson and Manning, 2019; Wang and Ling, 2025). Furthermore, VLA models frequently rely on external planners or heuristics, obscuring the locus of spatial competence (Chen et al., 2024; Song et al., 2025a). As a result, current VLA systems exhibit impressive surface behavior but limited internal spatial understanding (Pothiraj et al., 2025; Danish et al., 2025).

★ **Planning, Memory, and World Interaction.** Effective embodied spatial reasoning depends critically on memory and planning mechanisms capable of supporting long-horizon interaction. Many agents rely on short-term visual context or reactive policies, which degrade rapidly as environments grow larger or more complex (Zhang et al., 2025e; Song et al., 2025a). Persistent spatial memory, such as map-like representations, remains difficult to learn and maintain under noisy observations (Pothiraj et al., 2025; Danish et al., 2025). Planning failures often arise from inconsistent internal world models, causing agents to revisit explored areas or misjudge distances (Chen et al., 2024; Munje et al., 2025). These issues highlight the need for architectures that explicitly represent and update spatial state over time (Song et al., 2025a; Pothiraj et al., 2025).

H More Details on Failure Modes and Open Problems

Despite advances in scale and benchmarks, spatial reasoning in foundation models remains fragile and inconsistent, with recurring failures across tasks and architectures indicating fundamental limitations (Pothiraj et al., 2025; Wang and Ling, 2025). Here, we summarize key failure modes and open problems critical for robust spatial intelligence, and provide a structured mapping to mitigation strategies.

Table 3: Failure modes mapped to mitigation strategies and residual gaps.

Failure Mode	Approaches That Partially Mitigate	Where Gaps Remain
Shortcut learning	Synthetic diagnostics, counterfactual augmentation, scene-graph supervision (Suhr et al., 2017; Wang and Ling, 2025)	Real-world viewpoint shifts, rare configurations
Metric deficiency	Depth/point-cloud supervision, metric losses, 3D lifting (Chen et al., 2024; Danish et al., 2025)	Domain shifts, scalability, embodied settings
Perspective errors	Multi-view contrastive training, geometry-aware encoding (Danish et al., 2025; Munje et al., 2025)	Novel viewpoints, chained rotations
Knowing–doing gap	VLA models, RL with spatial rewards, tool augmentation (Zhang et al., 2025e; Song et al., 2025a)	Long-horizon consistency, sensor noise, replanning

♦ **2D Shortcut Learning and Dataset Bias.** A dominant failure mode arises from shortcut learning induced by biases in 2D image datasets. Models frequently exploit correlations between spatial relations and superficial visual cues, bypassing genuine relational reasoning (Johnson et al., 2017; Wang and Ling, 2025). Such shortcuts yield high benchmark accuracy while masking severe generalization failures under viewpoint change or counterfactual perturbation (Pothiraj et al., 2025; Suhr et al.,

2017). Synthetic diagnostics repeatedly demonstrate that models fail when these correlations are broken (Hudson and Manning, 2019; Tasnim et al., 2026). Addressing this issue requires both dataset redesign and evaluation protocols that explicitly penalize shortcut exploitation (Danish et al., 2025; Hudson and Manning, 2019).

♦ **Metric and Scale Failure.** Metric reasoning remains one of the weakest aspects of current spatial models, affecting distance estimation, size comparison, and volumetric understanding. Even models augmented with depth cues or 3D supervision often fail to maintain consistent internal scale across scenes (Chen et al., 2024; Danish et al., 2025). Errors in metric reasoning propagate to downstream tasks such as navigation and manipulation, where small inaccuracies compound into task failure (Song et al., 2025a; Pothiraj et al., 2025). Benchmark results indicate that metric failures persist even when qualitative relations are correctly identified (Munje et al., 2025; Hudson and Manning, 2019). Developing representations that enforce metric consistency remains a central open challenge (Chen et al., 2024; Pothiraj et al., 2025).

♦ **Perspective and Rotation Errors.** Perspective-dependent reasoning exposes systematic weaknesses in mental rotation and frame-of-reference alignment. Models frequently confuse egocentric, allocentric, and object-centric frames, leading to inconsistent answers across viewpoints (Stogiannidis et al., 2025; Pothiraj et al., 2025). These errors are particularly pronounced in multi-view and 3D benchmarks, where internal representations fail to align observations geometrically (Danish et al., 2025; Munje et al., 2025). Prompting and reasoning-time interventions offer limited mitigation, suggesting that the underlying representations remain deficient (Jia et al., 2026a; Wang and Ling, 2025). Robust perspective reasoning likely requires explicit mechanisms for transformation and invariance (Chen et al., 2024; Pothiraj et al., 2025).

♦ **The Knowing–Doing Gap in Agents.** A critical open problem is the knowing–doing gap, where agents can verbally describe correct spatial relations but fail to act on them effectively. This disconnect highlights a separation between declarative spatial knowledge and procedural execution (Song et al., 2025a; Zhang et al., 2025e). Empirical studies show that agents often generate valid plans in language while executing spatially incoherent actions (Pothiraj et al., 2025; Munje et al., 2025). Tool augmentation and external planners can reduce this gap but shift competence outside the model (Chen et al., 2024; Song et al., 2025a). Closing the knowing–doing gap remains essential for deploying spatially competent agents in real-world environments (Pothiraj et al., 2025; Danish et al., 2025).

I Toward Genuine Spatial Intelligence

Achieving genuine spatial intelligence requires moving beyond incremental benchmark gains toward principled advances in data, architectures, evaluation, and application domains. The evidence surveyed in previous sections suggests that current models exhibit fragmented spatial competence, excelling in isolated settings while failing to generalize across tasks, modalities, and environments (Pothiraj et al., 2025; Wang and Ling, 2025). Progress therefore depends on rethinking foundational assumptions about how spatial knowledge is acquired, represented, and evaluated in large-scale models (Chen et al., 2024; Hudson and Manning, 2019). In this section, we outline key research directions that emerge from the limitations identified throughout the survey.

♦ **Data Requirements and Scaling Laws.** Scaling laws have driven remarkable progress in language and vision, yet their applicability to spatial reasoning remains unclear. Empirical results indicate that simply increasing data and model size yields diminishing returns for metric, perspective-dependent, and long-horizon spatial tasks (Pothiraj et al., 2025; Danish et al., 2025). Many large-scale datasets underrepresent rare but critical spatial configurations, reinforcing shortcut learning rather than robust geometric abstraction (Wang and Ling, 2025; Johnson et al., 2017). Moreover, spatial reasoning appears to benefit disproportionately from structured diversity and targeted supervision rather than sheer volume (Stogiannidis et al., 2025; Hudson and Manning, 2019). These observations suggest that future scaling efforts must prioritize data diversity, relational coverage, and environment interaction over passive data accumulation (Chen et al., 2024; Pothiraj et al., 2025).

♦ **Architectural Shifts Toward World Models.** A recurring theme across failure analyses is the absence of persistent, coherent internal world models in current foundation architectures. While implicit latent representations suffice for recognition and shallow reasoning, they struggle to support prediction, counterfactual reasoning, and long-horizon

planning (Pothiraj et al., 2025; Danish et al., 2025). Architectures that incorporate explicit state, dynamics, and transformation mechanisms show promise, particularly in embodied and agentic settings (Song et al., 2025a; Zhang et al., 2025e). However, existing world-model approaches often prioritize visual realism over geometric and physical consistency (Chen et al., 2024; Wang and Ling, 2025). Future architectural shifts must therefore emphasize structured spatial state, controllable simulation, and integration with language-based abstraction (Munje et al., 2025; Song et al., 2025a).

♦ **Next-Generation Benchmarks.** Benchmark design will play a decisive role in shaping the trajectory of spatial reasoning research. The survey highlights that many current benchmarks overemphasize static accuracy while under-testing generalization, counterfactual robustness, and temporal consistency (Hudson and Manning, 2019; Pothiraj et al., 2025). Next-generation benchmarks should explicitly align with the framework introduced in Section 3, ensuring balanced coverage of relation types, reasoning processes, and representation demands (Danish et al., 2025; Wang and Ling, 2025). Incorporating controlled perturbations, multi-view consistency checks, and agentic interaction can help distinguish genuine spatial understanding from shortcut exploitation (Stogiannidis et al., 2025; Song et al., 2025a). Without such benchmarks, progress risks remaining benchmark-centric rather than capability-centric (Pothiraj et al., 2025; Hudson and Manning, 2019).

♦ **Implications for Robotics and Science AI.** Robust spatial reasoning is foundational not only for robotics, but also for scientific domains that rely on spatial structure, such as molecular modeling, climate simulation, and systems biology. In robotics, failures in metric reasoning, memory, and perspective alignment directly translate into unsafe or inefficient behavior (Song et al., 2025a; Zhang et al., 2025e; Spranger et al., 2014). Similarly, scientific AI applications require models that can reason about structure, scale, and transformation beyond surface-level pattern recognition (Chen et al., 2024; Munje et al., 2025). The limitations identified in this survey suggest that deploying foundation models in these domains without explicit spatial grounding remains risky (Pothiraj et al., 2025; Danish et al., 2025). Addressing these challenges offers the potential for transformative impact, but will require coordinated advances across data, architecture, evaluation, and theory (Hudson and Manning, 2019; Song et al., 2025a).