

Towards a Data-Centric Understanding of Spatial Reasoning in Multimodal Large Language Models

Mahdieh Mallahnezhad^{1*}, Masoud Jafaripour¹,
Mohamadali Tofigh¹, Sheila Schoepp¹, Osmar R. Zaiane^{1,2}

¹Department of Computing Science, University of Alberta, Edmonton,
Alberta, Canada.

²Alberta Machine Intelligence Institute (Amii) Fellow, CIFAR Fellow.

*Corresponding author: Mahdieh Mallahnezhad, mallahne@ualberta.ca;

Contributing authors: jafaripo@ualberta.ca; tofigh@ualberta.ca;

sschoepp@ualberta.ca; zaiane@ualberta.ca;

Abstract

Multimodal Large Language Models (MLLMs) have been widely applied in embodied agents and robotics for vision tasks. However, MLLMs suffer from limitations in spatial reasoning and perception. Post-training techniques have shown promise for improving these capabilities; yet, there are limited data-driven surveys of such approaches. In this survey, we consider approaches to fine-tuning pretrained MLLMs for spatial reasoning tasks, ranging from supervised to reinforcement learning, with an emphasis on the datasets used. We investigate themes of dataset construction, data modality and annotation, and evaluation of the resulting models. Based on these themes, we categorize research into three groups: dataset design, dataset construction and evaluation. We find four main methods for constructing fixed spatial annotations: rules and templates, LLM prompting, procedural generation, and human-in-the-loop. We also observe substantial variation in evaluation approaches, including different scoring methods, tolerance metrics, and benchmark types. Finally, we identify opportunities for future research in unified dynamic annotation, efficient spatial reasoning, standardization of evaluation, expanded spatial range, and human-centered spatial interactions.

Keywords: Multimodal large language models, Spatial reasoning, Training datasets, Data annotation, Survey

1 Introduction

Vision-Language Models (VLMs) and, more broadly, Multimodal Large Language Models (MLLMs) combine visual and language reasoning to interpret visual inputs, understand spatial and semantic relations between objects in those inputs, and produce answers to natural language questions based on them (Liu et al. 2023). Although these models perform well on image captioning, visual question answering, and object identification (Yin et al. 2024), they struggle with spatial reasoning (Cheng et al. 2024). For example, they often fail when asked questions about spatial relations that should be relatively easy to answer, such as relative positions (“Is the cup to the left of the plate?”), distances (“How far is the table from the wall?”), size comparisons (“Which object is bigger?”), and even different view angles (“What would this scene look like from behind?”) (Chen et al. 2024; Huang et al. 2026). This presents a significant problem: even though MLLMs can analyze and describe objects within a scene, they can make incorrect inferences about those objects’ locations or their spatial relationships with each other.

Spatial understanding is the ability to make inferences about spatial layouts and spatially relevant properties of objects. Humans use it in everyday activities: to navigate around a crowded room, to place an object on a shelf, or to decide whether a box fits in a particular location (Vieu 1997). It is not just a perceptual ability; spatial understanding plays an essential role in communicative interactions, action planning, and reasoning about the physical world (Tversky 2005).

Spatial reasoning has accordingly become a central challenge for embodied artificial intelligence (AI) (Liu et al. 2026a). An embodied agent must be able to perceive, plan, and act in the physical world. Without a firm grasp of spatial relations, agents will struggle to map perceptions to the corresponding actions (Yuan et al. 2024). Spatial understanding therefore acts as a bottleneck on the deployment of embodied VLM/MLLM agents in real-world environments (Song et al. 2025).

A number of mechanisms allow MLLMs to complete tasks involving spatial reasoning. The model architecture determines how visual and linguistic information are represented; e.g., how RGB data are combined with depth data can have a significant effect on a model’s spatial reasoning abilities (Cai et al. 2025; Liu et al. 2025b). The training method also affects spatial reasoning outcomes; e.g., post-training with supervised fine-tuning (SFT) and reinforcement learning (RL) using verifiable rewards can have implications for how the models produce outcomes (Ouyang et al. 2025; Batra et al. 2025).

Another perspective suggests that it is the training corpus that is a key factor limiting MLLMs’ spatial reasoning abilities. In Chen et al. (2024), the authors hypothesize that the limitations in current VLMs are not due to the model architecture itself, but due to the lack of spatial information present in the training data. This hypothesis is consistent with Ogezi and Shi (2025), who found that spatial relations are underrepresented in the datasets they examined and that the performance of VLMs can be improved by introducing synthetic spatial relations. While these studies do support the importance of spatial supervision, they do not eliminate the potential influence of architecture, training objective, or the general-purpose nature of the training data. These results encourage research that focuses on the data used for training

and the creation of datasets for spatial reasoning, employing strategies that range from procedurally-generated synthetic scenes to large-scale automatic annotation of real-world images (Deng et al. 2026; Cheng et al. 2024; Song et al. 2025).

Recent surveys of MLLMs have explored the area of spatial reasoning from either an architecture- and training-focused (Zheng et al. 2025) or task and benchmark perspective (Liu et al. 2025a). There are also data-centric surveys that explore the creation and utilization of multimodal training datasets (Bai et al. 2024). There are also broader surveys that look at data-driven methods and spatially trained datasets (see Section 2.2). Our survey adds to these by looking at dataset creation pipelines, annotation sources, and comparison of evaluation methods as our primary focus. Since the data sources and their design appear to be one of the major factors limiting current models’ performance on these tasks (Chen et al. 2024; Deng et al. 2026), such a survey can help guide future research in this area. We do so by reviewing datasets used to fine-tune pre-trained foundation models for spatial reasoning and the evaluation of the models produced. We use the term fine-tuning broadly to refer to any update to model parameters/structure after pretraining, including supervised instruction tuning, parameter-efficient tuning (e.g., PEFT/LoRA), preference tuning, and RL with spatially grounded rewards (Li et al. 2026; Shen et al. 2025; Wu et al. 2025).

We focus on selected training datasets and pipelines that produced them, and we discuss benchmark datasets only when they accompany a training dataset, in order to contrast learning- versus evaluation-focused design choices. Table 1 summarizes the datasets we review. In short, the contributions of this survey are:

- A new taxonomy of selected datasets for MLLM fine-tuning for spatial reasoning (Figure 1).
- A critical analysis of the dataset construction pipelines, and how synthetic and human annotations are blended to produce training datasets.
- An audit of evaluation protocols across the surveyed datasets, revealing inconsistent scoring criteria, tolerance thresholds, and benchmark construction that limit cross-work comparability.
- A proposal outlining insights, gaps, and future directions, including unexplored spatial tasks, scalability issues, and cross-modal and cross-domain generalization.

The rest of this survey follows the three branches of the taxonomy in Figure 1. Section 2 provides background context and reviews related surveys. Section 3 examines dimensions of dataset design, including ground truth, task patterns, viewpoints, and reference frames. Section 4 covers dimensions of dataset creation, including construction methods, annotation approaches, modality, and source environments. Section 5 addresses aspects of evaluation, including what is evaluated, how answers are scored, and how generalization is evaluated. Section 6 makes comparisons across datasets and identifies patterns, open problems, and opportunities for future work.



Fig. 1 Taxonomy of data-centric spatial reasoning in MLLMs, organized by dataset design, construction, and evaluation.

2 Background

2.1 Spatial Awareness and Visual Reasoning

Spatial awareness refers to the perceptual and cognitive abilities to derive spatial information from visual input and to maintain an internal representation of the spatial configuration of objects in a scene (such as their location or relative positioning) [Ranasinghe et al. \(2024\)](#). The presence of an internal representation that is independent of the observer has been proposed across many studies in cognitive science [O’Keefe and Nadel \(1979\)](#). Internal representations have been linked to spatial manipulation and mental simulation tasks, which we refer to as spatial cognition [Hegarty \(2004\)](#). While these two terms are often used interchangeably in the literature, we use the distinction here to help organize the survey of these capabilities, as some works use spatial cognition to refer to what we classify as spatial awareness [Cheng et al. \(2024\)](#); [Zhao et al. \(2025\)](#). This forms a foundational capability for visual reasoning and interaction with the environment [Piaget and Inhelder \(2013\)](#).

Visual reasoning is commonly assumed to build on spatial awareness to perform high-level reasoning over visual input. It uses perceptual and spatially aware knowledge to draw inferences, make comparisons, and answer questions that go beyond observing visual scenes. Prior work has shown a deficiency in language-only reasoning for spatial reasoning [Li et al. \(2024a\)](#). However, merely reasoning over visual information is not sufficient to overcome these shortcomings. There are still many failures in spatial relations in vision-language models [Kamath et al. \(2023\)](#). These shortcomings emphasize the need for specialized post-training and fine-tuning, which is the focus of this survey, in order to correctly align MLLMs with spatial concepts.

2.2 Related Surveys

Recent surveys have investigated spatial reasoning in MLLMs from complementary angles. From a general multimodal perspective, [Bai et al. \(2024\)](#) offer a dataset-level, modality-agnostic survey of datasets and benchmarks for evaluating MLLMs, though not specifically with respect to spatial reasoning. From the narrower focus of spatial reasoning, [Liu et al. \(2025a\)](#) survey spatial reasoning tasks, benchmarks, and methods from a cognitive science view: they propose a taxonomy based on cognitive complexity, and map existing benchmarks to it, as well as model architectures and algorithms, and survey how they are (or can be) evaluated and improved. [Zheng et al. \(2025\)](#) focus on model architectures, capabilities, and post-training extending the discussion from 2D VLMs to 3D and embodied AI environments, and propose benchmarks for evaluating spatial reasoning. [Feng et al. \(2025\)](#) broaden the scope along a different axis, surveying LLM-powered spatial intelligence across embodied, urban, and earth scales through the lens of human spatial cognition. [Ma et al. \(2024\)](#) focus on 3D settings and survey approaches to connect LLMs to 3D information, as well as provide a meta-analysis on the current state of 3D-LLMs. Similarly, [Zha et al. \(2025\)](#) survey methods for enabling LLMs to understand 3D space, categorizing these approaches into image-based, point cloud-based, and hybrid methods. In addition to these method-focused surveys, [Yu et al. \(2025\)](#) survey visual spatial understanding in VLMs via

benchmarks. They consolidate nearly 20 open-source datasets into a benchmark suite (SIBench) that highlights a large disconnect between visual perception and spatial understanding. Lastly, [Liu et al. \(2026a\)](#) provide a comprehensive survey of spatial understanding in VLMs. They cover definitions of spatial tasks, error analysis, methods for improvement, evaluation techniques, and use cases. Furthermore, they propose a taxonomy based on intervention methods and support this with an extensive empirical study. To complement these works, our survey focuses specifically on the construction and comparison of selected spatial post-training datasets, their annotation schemes, and how model performance is evaluated.

3 Spatial Reasoning Dataset Design Dimensions

Datasets used for fine-tuning foundation models to improve spatial understanding typically include paired visual inputs with textual annotations encoding different spatial attributes such as object relations, depth and distances, geometric structures, and layouts [Liu et al. \(2026a\)](#). When the annotations are in the form of question-answer pairs, they can be used to create a visual question answering (VQA) dataset that is commonly used for fine-tuning. In supervised instruction tuning, these models are fine-tuned as conditional text generators given visual and textual inputs for next-token prediction with multimodal supervision, so VQA is a model-agnostic format for fine-tuning datasets ([Liu et al. 2023](#); [Li et al. 2023](#)). However, some datasets extend beyond conventional VQA by feeding different modalities to the model or encoding their cues into textual annotations (see Section 4 for more details).

3.1 Ground Truth Spatial Reasoning

While VLMs are typically pre-trained on large-scale, unstructured multimodal corpora through self-supervised objectives, their post-training effectiveness depends heavily on the quality of ground-truth annotated datasets. Ground truth in machine learning, in general, is an external reference signal for supervising the learning process [Jaton \(2021\)](#), and in LLM reasoning, it is a set of external signals used to evaluate the reasoning trace [Lee and Hockenmaier \(2025\)](#). In spatial reasoning, this definition can be extended to a set of structured reference specifications of spatial information, such as object relations and locations, depth, and compositional plans, used to train or evaluate a model’s performance given a visual or multimodal input [Johnson et al. \(2017\)](#). From a data-centric perspective, building ground-truth spatial understanding is challenging since most spatial attributes are implicit, and their availability and reliability depend on the source of information. Therefore, it has to be created through a carefully designed dataset generation pipeline, either autonomously or under human supervision.

3.2 Patterns of Tasks: Perception to Reasoning

To train VLMs for spatial understanding, datasets derive a spatial learning signal from a ground truth source. This is typically done in practice by defining task types that specify what the spatial reasoning is meant to teach. To provide a view of some

recurring patterns, at the simplest level, many datasets define a task that relates to perception, such that the model’s reasoning is expected to relate that spatial language to visual evidence, such as identification, localization, or presence of an object. SpatialLadder Li et al. (2026) defines this approach in the context of a hierarchy of tasks, with perceptual reasoning defined as the first level of a hierarchical curriculum, claiming that higher-level forms of reasoning can only be reliably performed if perception is accurate. Relational reasoning tasks require the model to reason about the relationships between objects in the image, rather than relying on perception alone. This typically takes the form of pairs of objects in the scene, which are annotated with qualitative relationships (e.g., left-of, above, behind) or comparative relationships (e.g., closer to, larger than). In this case, ground truth is typically derived from 3D scene metadata or depth of objects in the scene Cheng et al. (2024). A variation on this task type is perspective-transformation reasoning, which requires reasoning in relation to different perspectives, such as the camera’s motion or matching objects in images with different camera poses Deng et al. (2026). Temporal and dynamic reasoning builds upon this with sequences: one can use static relations to reason about something being “above” another, but dynamic relations such as “into” or “onto” require motion in their creation. As a result, ground truth for dynamic reasoning can be obtained from video sources Ouyang et al. (2025); Wang et al. (2025). However, dynamic reasoning is not exclusively obtained from video: Spatial-DISE also uses dynamic reasoning on objects from static images (Huang et al. 2026). Thus, motion of objects, motion of the camera, and a hypothetical transformation should be clearly specified for dynamic reasoning tasks.

Metric reasoning requires models to estimate quantities, such as absolute distances or object sizes. Ground truth is obtained either from depth sensors Cai et al. (2025) or from depth maps estimated during data generation Cheng et al. (2024); Chen et al. (2024). Less common tasks include occlusion-aware reasoning from visibility masks Lee et al. (2025) and planning tasks that indicate action sequences to achieve spatial objectives Sun et al. (2026). These task categories typically fall into one of the splits defined in Section 2. For perceptual and metric tasks, the model must use the input to gain knowledge about spatial relationships, or possess spatial awareness. For perspective-transformation, dynamic, occlusion-aware and planning tasks, the model must manipulate or extrapolate from an already-existing representation, or possess spatial cognition. Relational reasoning straddles this divide as reading a relationship from an observed pair can be considered perceptual, but comparing across objects in a transformation cannot. Most datasets include examples from both sides of this divide.

3.3 Spatial Viewpoints and Frames of Reference

Beyond task selection, spatial reasoning datasets also need to define a viewpoint for observation. Many use a single-view approach, with datasets like SpatialVLM Chen et al. (2024), and SpatialRGPT Cheng et al. (2024) testing whether a model can infer a 3D relation, for example, a mug being behind a laptop, from a singular observation of a 2D scene. On the other hand, InternSpatial Deng et al. (2026) builds off this notion of single-view and multi-view, providing spatial supervision such that models need to

connect information from multiple observations of the same spatial scene rather than relying on just one view.

The way spatial information is observed influences how it is represented and also the choice of reference frame influences what spatial relations can even be expressed. In general, spatial relations are defined relative to a reference frame. In an egocentric reference frame, relations between objects are expressed relative to the observer. This is especially relevant to embodied agents [Zhao et al. \(2025\)](#). In a world-centric reference frame, relations between objects are expressed in world coordinates external to the observer. For example, an object can be expressed as being above another object irrespective of the location of the observer in the world [Chen et al. \(2024\)](#). In an object-centric reference frame, spatial information is expressed relative to the object itself. This distinction of frames has an influence on correct answers; for example, in RoboSpatial [Song et al. \(2025\)](#), a bowl can be in front of a car from a third-person perspective but behind the car in an egocentric perspective. Thus, for the same scene, different answers can be correct depending on the reference frame. Finally, some datasets employ a flat 2D map view of the environment, as in SpaceR-151k [Ouyang et al. \(2025\)](#).

Spatial reasoning datasets that ask the models to reason about spatial relations in multiple frames of reference, such as SenseNova-SI-8M [Cai et al. \(2026\)](#), help the models to develop an understanding of how the same spatial relation can be described differently depending on the chosen frame of reference.

Table 1 records the viewpoints, reference frames, and spatial representation of each selected dataset: single-view settings dominate compared to multi-view settings. Egocentric framing is the most common frame, and four datasets adopt a top-down representation.

4 Dataset Building

This section examines how spatial reasoning datasets for VLMs are built in practice. We review how datasets are constructed or adapted from existing data resources, how spatial information is annotated, which modalities are supported, and which environments serve as the sources of data. Most of these datasets use a similar data pipeline as shown in Figure 2. Raw data may include single RGB images, video or multi-view images, or 3D simulations. Using perception and geometry models these inputs are transformed into depth, keypoints, bounding boxes and segmentation masks. In other pipelines, geometry can be derived from captured 3D data or the states of a simulator. Spatial knowledge can also come from human annotations or captions. Then the masks, visual labels and maps that are extracted, are passed into a more structured spatial representation. At this step, the information undergoes additional transformations to add semantic, spatial and 3D information by having organizing geometry, camera parameters, coordinates, relationships, object instances and spatial information. Now, the spatial supervision signal is ready to be used by annotation methods. The annotation stage then turns this signal into training data through templates, LLM-generated rationales and captions, human supervision, procedural computation, or by the model-generated labels. Annotation produces VQA pairs and, in some cases,

Table 1 Spatial reasoning post-training datasets and dataset collections. Several entries also serve as benchmarks; these are included where a training split accompanies the benchmark. Codes are defined in the legend below.

Citation	Name	Scale	View	Modality	Annotation
Song et al. (2025)	RoboSpatial	4M ^Σ	single, object, ego, world	rgb, bbox, dmap, ptcl, cpose, nl	lm, proc, tmpl
Chen et al. (2025a)	Robo2VLM-1	685k	multi, ego, world	rgb, seg, nl, traj	tmpl, semi
Ning et al. (2025)	3D ReasonSeg	29.2k	world	seg, ptcl, nl	lm, tmpl, semi
Liu et al. (2025b)	SSR-COT	1.2M	single, world	rgb, dmap, nl	lm, tmpl, other
Deng et al. (2026)	InternSpatial	12M	single, multi, ego	rgb, bbox, seg, nl	lm, proc, tmpl, semi
Ranasinghe et al. (2024)	Localize [†]	200k+	single	vid, rgb, bbox, nl	lm, tmpl, semi
Zhang et al. (2025b)	Open3D-VQA	74.2k ^Σ	single, topdown, ego, world	object, rgb, bbox, seg, dmap, ptcl, cpose, nl	lm, proc, tmpl, semi
Yuan et al. (2024)	RoboPoint-1432K	1432k	single, ego, world	rgb, bbox, kpt, dmap, nl	proc, tmpl
Ogezi and Shi (2025)	SpaRE	3.8M ^Σ	single, object	rgb, nl	lm, proc, semi
Tang et al. (2025)	Sparkle	180k ^Σ	single, topdown, world	rgb, nl	proc, tmpl
Huang et al. (2026)	Spatial-DISE-12K	12.4k	single, multi, object	rgb, nl	proc, tmpl, semi
Qiyang et al. (2025)	SpaceR-151k	151k	single, topdown, ego	vid, rgb, bbox, ptcl, nl	proc, tmpl
Batra et al. (2025)	STVQA-7K	7.6k	single, ego	rgb, bbox, nl	lm, full
Shen et al. (2025)	Open Spatial Reasoning	900k ^Σ	single, object	rgb, seg, dmap, nl	lm, proc
Wu et al. (2025)	VILASR [†]	80.8k ^Σ	single, multi	vid, rgb, bbox, nl	lm, proc, tmpl
Sarch et al. (2025)	ViGoRL	86.6k ^Σ	single, object	rgb, nl	lm, other
Chen et al. (2025b)	—	409k ^Σ	single, ego	rgb, nl	lm, tmpl, semi, full
Zhao et al. (2025)	—	5.4k	ego, world	vid, rgb, kpt, nl, traj	—
Chen et al. (2026)	—	247k ^Σ	multi, ego	rgb, nl	lm
Cai et al. (2025)	SpatialQA [†]	746k	single, object, ego	rgb, bbox, dmap, nl, traj	lm, proc, tmpl, semi
Li et al. (2026)	SpatialLadder-26K	26.6k	single, multi, ego, world	vid, rgb, bbox, nl	proc, tmpl
Cheng et al. (2024)	OSD	8.7M	single, ego, world	rgb, bbox, seg, dmap, nl	lm, proc, tmpl
Chen et al. (2024)	—	2B	single, ego, world	rgb, bbox, seg, dmap, ptcl, nl	proc, tmpl
Wang et al. (2025)	STV-205k	205k	multi	vid, rgb, bbox, nl	proc, tmpl, full
Sun et al. (2026)	SpaceVista-1M	1M	multi, ego, world	vid, rgb, bbox, kpt, seg, nl	lm, proc, tmpl, semi
Zhang et al. (2025a)	SPAR-7M	7M	single, multi, ego	vid, rgb, bbox, kpt	lm, proc, tmpl
Lee et al. (2025)	SpatialMosaic	2M	multi, ego	rgb, bbox, dmap, ptcl, cpose, nl	proc, tmpl
Cai et al. (2026)	SenseNova-SI-8M	8.5M	single, multi, topdown, object, ego, world	vid, rgb, bbox, ptcl, nl	proc, other

Name. [†] marks dataset families reported under a single name: Localize (Localize-Instruct-200K, Localize-ActivityNet); VILASR (VILASR-ColdStart-33k, VILASR-RRS-8k, VILASR-RL-40k); SpatialQA (SpatialQA, SpatialQA-E). A dash (—) denotes not reported, here and in every column.

Scale. Rounded QA/example counts unless marked; $B/M/k$: billions/millions/thousands; $+$: lower bound. Superscript Σ : survey-computed sum across visual/annotation records or stage collections.

View. Combines viewpoint with frame of reference. Viewpoint: single-view (*single*), multi-view (*multi*). Frame: object-centric (*object*), egocentric (*ego*), world-centric (*world*). Top-down (*topdown*) denotes a flat 2D map representation.

Modality. Vision: video (*vid*), RGB (*rgb*), bounding box (*bbox*), keypoint (*kpt*), segmentation (*seg*). Geometry: depth map (*dmap*), point cloud (*ptcl*), camera pose (*cpose*). Text: natural language (*nl*), trajectory (*traj*).

Annotation. How the fixed training corpus was annotated. Automated: LLM-based (*lm*), procedural or 3D reconstruction (*proc*), template- or rule-based (*tmpl*), supervision derived from auxiliary vision models or from the model's own outputs (*other*). Human involvement: partial (*semi*), full (*full*).

reasoning traces. The following subsections examine each of these stages with more details.

4.1 Dataset Construction Strategies

Most of the surveyed spatial reasoning datasets do not build their visual data collection from the ground up, but instead extend general-purpose vision datasets, reformatting their annotations into spatial QA style, and allowing for spatial supervision to be extended to existing datasets and scaled without new data collection. For instance, [Ranasinghe et al. \(2024\)](#) convert bounding boxes to instruction-based QA with coordinate answers and scale with automated data generation. Other work extends even further, building on existing 2D or 3D datasets and extending them with automatically computed spatial annotations (e.g., robot-relevant signals such as free space), and then generating a massive scale of spatial QA pairs from those [Song et al. \(2025\)](#). Some approaches leverage multiple datasets to maximize coverage and diversity. For instance, SpaceVista-1M mixes sources in their curated dataset, combining indoor scans, outdoor scenes, and aerial videos [Sun et al. \(2026\)](#). Similarly, SPAR-7M [Zhang et al. \(2025a\)](#) reformats existing richly annotated 3D indoor scene datasets into 7M QA pairs spanning 33 task types across single-view, multi-view, and video inputs, using scene geometry to generate supervision that ranges from 2D perception to 3D-grounded reasoning. SenseNova-SI [Cai et al. \(2026\)](#) takes this aggregation strategy further by combining 4M samples from different open-source datasets under a unified spatial-capability taxonomy, then generating another 4.5M QA pairs from richly annotated 3D sources (ScanNet, ScanNet++, SUN RGB-D, Matterport3D, and others) to cover less represented capabilities such as perspective-taking, for a total of 8.5M pairs. Few of the recent datasets are built from scratch, which are typically designed for a specific target task or for benchmarking rather than training at scale. Sparkle ([Tang et al. 2025](#)), for instance, generates a small-scale synthetic dataset that teaches basic spatial reasoning capabilities (such as direction, distance, and localization) with the goal of generalizing to harder tasks. Spatial-DISE [Huang et al. \(2026\)](#) has a similar focus on simple abilities, but is mainly designed as a hybrid benchmark dataset, combining real-world problems with synthetic scenes.

4.2 Annotation Approaches

The annotation approach used to create a fine-tuning dataset inherently constrains what spatial concepts a VLM dataset can directly supervise. The annotation of the training data for spatial understanding, whether as coordinate pairs, natural language text, reasoning paths, or the outputs of a model, exposes models to the “ground truth” spatial structure that may be underrepresented in pre-training. Additionally, annotation approaches vary in their quality, but all come to a quality-quantity equilibrium with varying scale of supervision, precision, coverage, and human involvement. We examine annotation approaches that lead to the creation of fine-tuning datasets, with a spectrum from automated to human. Across most annotation approaches, the pre-trained perception model extracts object- and geometry-level structure from the

visual input using auxiliary models (e.g., object detectors, depth or geometry estimators), which downstream templates, LLMs, procedural rules, or learning objectives then leverage to generate spatial supervisions (Chen et al. 2024; Cheng et al. 2024; Song et al. 2025). The Annotation column of Table 1 lists the annotation method for each dataset. These annotation methods represent potential elements that can appear within the same pipeline and are not necessarily mutually exclusive. Instead, they can be combined at various stages of annotation.

4.2.1 Template-Based and Rule-Based Annotation

In template-based annotation, a rule and a sentence are paired together. The rule is applied to a structured scene and produces an answer to a relation (e.g., the location of one object relative to another or the distance between two objects). The template then expresses that relation as a question. Because the mapping goes from relation to text, the answer is generated prior to the question and a single rule can be applied across an entire corpus (see Figure 3). SpatialVLM (Chen et al. 2024) defines 38 types of spatial questions that can express either qualitative or quantitative relations. Each question and answer type is expressed using a small number of templates to create a large and balanced corpus of questions and answers. Similarly, SpatialRGPT (Cheng et al. 2024) converts 3D scene graphs into region-level questions and answers. Hand-crafted templates are used to ask questions regarding object dimensions (height/width) and qualitative relations (left/right or direction), as well as quantitative metrics (distance). While this system uses a mix of approaches, these templates are the foundation of its large collection of spatial questions and answers. InternSpatial (Deng et al. 2026) develops an automated system that converts varied visual inputs into spatial questions and answers. Each question is then converted into visual and textual references using template-based methods that utilize 3D view space. SpatialLadder (Li et al. 2026) enforces consistency across questions and answers by using a small number of question templates across modalities and tasks. Geometric consistency is maintained by leveraging 3D reconstructions for template instantiations. Sparkle (Tang et al. 2025) decomposes 2D spatial reasoning into three components: direction, distance, and localization, and designs prompt templates for each, showing that models trained on these basic templates generalize to other composite tasks (e.g., shortest path problems).

One drawback of template-based approaches is linguistic inflexibility. InternSpatial (Deng et al. 2026) identifies limitations and addresses them by introducing LLMs to paraphrase and expand template-generated questions, before the QA pairs are instantiated, without losing their spatial semantics. However, this issue is not equally problematic across all scenarios. When using GPT-4 to rephrase the evaluation questions and answers in SpatialRGPT-Bench, SpatialRGPT-7B’s qualitative score (concerning rephrasing procedure) only drops by 1.8 points (from 91.8 to 90.0) (Cheng et al. 2024). A more general issue with templates is that they can only provide supervision data for spatial relationships that are present in the source representation. This design choice marks a natural transition to annotation pipelines in which LLMs play a generative role.

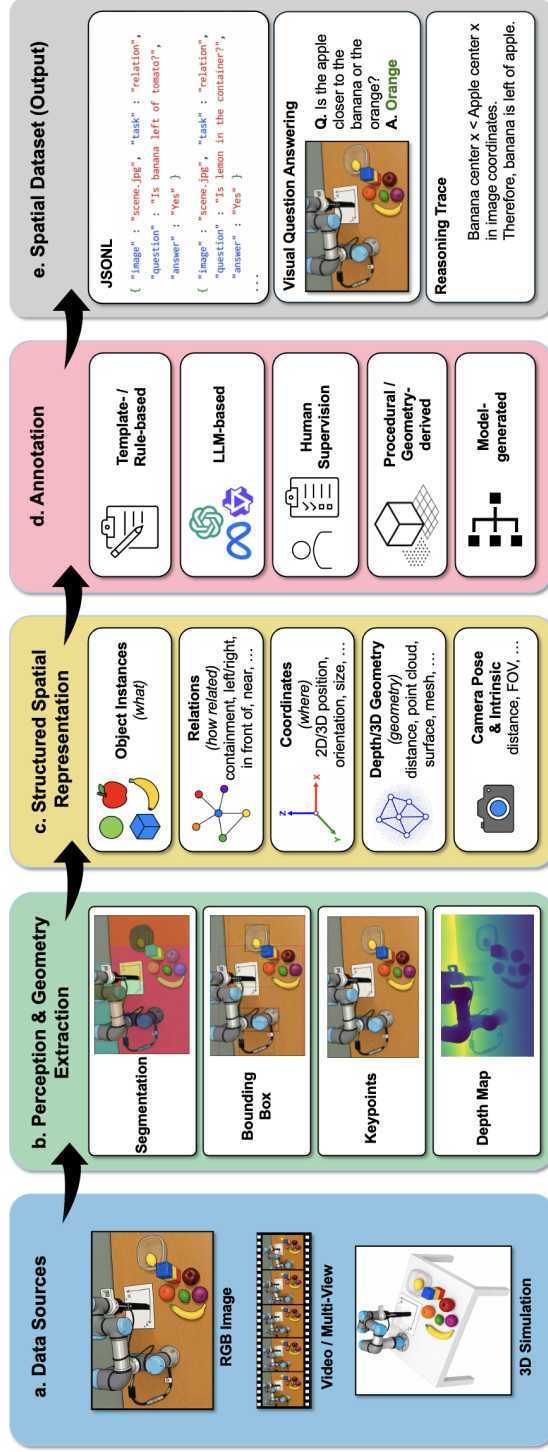


Fig. 2 Overview of the common pipeline across the literature for building spatial reasoning datasets. Where needed, raw visual data are first processed by perception and geometry models to extract spatial cues. Then they are fed into structured spatial representations. These representations provide the supervision signal used by different annotation approaches to generate the final spatial training data, which includes VQA pairs and optional reasoning traces.

Spatial VQA Templates (.jsonl / .yaml)		
 Scene Objects [A], [B], [C] = Objects (e.g., banana, tomato, lime, lemon, apple, onion, container, cubes, robot end-effector). [x, y] = coordinate (e.g., at robot base).	Task Type	Question Template
	Relation	Is [A] to the left of [B]?
	Direction	Where is [A] relative to [B]?
	Distance	How far is [A] from [B]?
	Comparison	Which is closer to [C], [A] or [B]?
	Localization	Where is [A] located from [x, y]?
	Size	Which object is larger, [A] or [B]?
	Containment	Is [A] inside [B]?
	3D Understanding	Which object is taller in 3D view [C], [A] or [B]?
	Answer	
		Yes
		Below
		10 cm
		Lime
		[-26, 34] cm
		Banana
		No
		Blue cube

Fig. 3 Template- and rule-based spatial VQA annotation (example from a robot-manipulation task). Explicit spatial rules define ground-truth and reference relations over structured scene representations, which are verbalized through controlled templates to produce scalable supervision. These are also conditional on precise object bindings and coordinate conventions.

4.2.2 LLM-Based Annotation

LLMs offer a flexible means of introducing linguistic variation into systems to create questions that are difficult to define with hand-written templates. Spatial-RGPT (Cheng et al. 2024) starts with facts from templates about spatial relations and uses a language model from the Llama family to create prompts. SpaRE (Ogezi and Shi 2025) begins with image captions and uses a Qwen language model to generate pairs of questions and answers, considering only those image captions that contain explicit information about a form of spatial reasoning.

LLM-annotation is useful for generating explanations, as well as step-by-step reasoning. 3D ReasonSeg Ning et al. (2025) considers the step-by-step reasoning generated by LLMs to be an explicit annotation, generating question-reasoning-answer triples all grounded in 3D spatial relations. SSR Liu et al. (2025b) instead uses the LLM-generated rationales as intermediate supervision, translating depth-derived spatial hints into a structured trace of reasoning that can be distilled into latent embeddings. In addition, LLMs can serve as automatic annotators and reviewers in the loop. In SpatialThinker (Batra et al. 2025) Claude Sonnet4 is used to generate question-answer pairs and to assign a difficulty rating to each question. Afterwards, GPT-4o is used for consistency checking between questions, answers and their spatial information. This is done with the goal of preserving well-grounded and reliable samples through multi-model consistency checking. A problem with using LLM-based annotation is the risk of hallucination or misinterpretation of the annotated visual content Liu et al. (2024). To reduce this issue, alternative datasets utilize procedural generation and 3D reconstruction techniques to create and annotate spatial relationships directly from geometry rather than language.

4.2.3 Procedural Generation and 3D Reconstruction

Procedural generation allows spatial reasoning datasets to have ground-truth knowledge for concepts where it is impractical to have humans label reliably at scale. The most impressive application of this is SpatialVLM’s (Chen et al. 2024) 3D lifting

pipeline, reconstructing estimated metric-scale point clouds from internet images using off-the-shelf depth and segmentation models. The reconstruction allows metric distance questions like “How far is object A from B in meters?” to be answered, which would otherwise be difficult for a human to do. RoboPoint (Yuan et al. 2024) combines procedural generation with an affordance prediction setting, producing spatial training data for robots from a 3D scene plus camera pose from a simulation, yet generalizes to real robots. RoboSpatial (Song et al. 2025) extracts robot-centric spatial annotations from real images and 3D scans, capturing reference frames and relations neglected by typical image captions. While these ideas depend on reconstructed or simulated 3D information to obtain spatial labels, SpatialMosaic (Lee et al. 2025) relies on it to capture a property of the view from high-quality ScanNet++ and Waymo Open Dataset point clouds. Using these point clouds, it then procedurally computes occlusion of objects across views and chooses pairs of views where the object is visible in the answer but occluded in the query. Based on these view pairs, it creates questions involving partial visibility in both indoor and outdoor environments. In summary, these methods automatically generate spatial annotations from reconstructed or simulated 3D data. However, they can only generate annotations for properties that can be computed from the available geometry. Errors in the estimated 3D data can be propagated into the generated annotations. For desired supervision that cannot be derived from the 3D data, other annotation methods can be used in conjunction with these approaches.

4.2.4 Model-Generated Supervision

The annotation schemes described in Table 1 fix questions, answers, or spatial annotations prior to training. In addition, the methods in this section also use model-generated outputs to create preference data before optimization, or to acquire reward signals during training. These approaches can utilize the annotations from the table but do not require annotated chains of reasoning for each example. Correctness signals on answers still need either reference answers or a defined verifier. SpatialThinker (Batra et al. 2025) incorporates scene graph creation into its reasoning output and then employs online reinforcement learning using rewards for answer correctness, number of regions, output formatting, and spatial relations grounding. SpaceR (Ouyang et al. 2025) employs spatially guided reinforcement learning for video reasoning tasks and uses task-specific verifiable rewards as well as a mapping reward that compares the predicted locations of objects to corresponding reference locations in a grid map. Embodied-R (Zhao et al. 2025) is applied to embodied tasks and provides rewards for correctness, logical consistency, and output formatting to encourage the reasoning to justify the final answer. Model-generated outputs can also be used as supervision via preference learning. Fine-Grained Preference Optimization (fDPO) (Shen et al. 2025) employs Multi-Model Monte Carlo Tree Search (M3CTS) to generate reasoning traces for supervised fine-tuning and preference pairs before performing preference optimization on offline-generated responses.

4.2.5 Human Supervision

Human supervision provides an intermediate level of automation, which incorporates a human touch at strategic points in the pipeline. Humans can ask insightful and well-justified spatial questions, but the millions of QA pairs used to train a VLM cannot be easily annotated manually. Some hybrid models therefore use human labor in earlier stages to guide the design or the generation process. Robo2VLM (Chen et al. 2025a), for example, had an expert review a small set of generated question-answer pairs before full dataset generation, using the feedback to refine the prompt and generation pipeline. Spatial-DISE (Huang et al. 2026) takes a more design-oriented approach, manually defining a cognitively grounded taxonomy of spatial reasoning tasks before deploying an automated pipeline to scale it up. Several other works build on existing resources that are annotated by human experts. For example, SpatialThinker (Batra et al. 2025) draws its spatial reasoning questions from the annotated ground truth of Visual Genome. Finally, when the goal is reliable evaluation, some introduce small, manually verified benchmarks along with their training data; RoboSpatial follows this approach with a compact hand-annotated evaluation set (Song et al. 2025).

4.3 Data Modalities

Many datasets share the same baseline: a sample of RGB images, questions, and answers. Various datasets include additional elements beyond this baseline: some elements are derived from the visual modality (e.g., bounding boxes or masks), which do not introduce a new sensor signal but allow for the division of an image into objects. Others provide geometry or sensing channels (for example, depth maps or point clouds), which can come from sensors, reconstruction, or estimation from RGB. Other additional elements are a temporal dimension to the visual data (for example, video sequences or trajectories of objects within those videos). Table 1 lists the data modalities of each dataset.

Vision and Language.

The simplest form of these datasets involves pairing the RGB images with language questions. VLMs are fine-tuned in these settings to act as conditional text generators, and VQA is a model-agnostic label type. A few datasets do, however, exclude vision altogether and instead ask the models to perform spatial reasoning based on language alone (Li et al. 2024b). Such settings help to determine whether the failure of VLMs to answer spatial language questions is due to failures in perception (vision) or reasoning (language) alone.

Bounding Boxes and Segmentation Masks.

Although bounding boxes, segmentation masks, and keypoints are often treated as inputs, they are a form of structure that can be provided by human annotators or computed from RGB images or geometry. These features are used to break apart the visual scene into separate, named objects. For instance, the language model might be asked to recognize that the mug is to the left of the laptop, which requires the model to understand the locations of both objects in the visual scene. Furthermore,

these features are often an output of the models rather than an input: models such as RoboPoint (Yuan et al. 2024) and Ranasinghe et al. (2024) explicitly encode these coordinates as text tokens to force the models to ground their language in the visual scene, while models such as SpatialThinker (Batra et al. 2025) use these coordinates as a target output to ensure the models can recognize the objects they are asked to reason about. Finally, SpatialLadder (Li et al. 2026) learns to recognize these objects as part of the first stage of its curriculum to enable later spatial reasoning abilities.

Depth Maps and Point Clouds.

While supervision from RGB-only images does not uniquely offer information regarding the metric distances within the image, which is essentially the lack of capability of VLMs (Cai et al. 2025; Chen et al. 2024), there are different ways of incorporating 3D information into the models. The geometry information may be incorporated either during the generation of the answer to the question, or it may be provided as an input to the model itself. For instance, SpatialVLM (Chen et al. 2024) utilizes monocular RGB images to reconstruct a point cloud through depth and camera pose estimation methods, using that 3D information to generate questions and answers that are used to supervise the learning of the model; the model is provided only with the RGB image and the question to be answered during inference. SpatialLadder (Li et al. 2026) utilizes the same approach for their models, but instead uses existing 3D reconstructions from ScanNet to obtain the 3D information. In InternSpatial (Deng et al. 2026), both methods are used based on the dataset used. For 3D datasets, existing geometry and camera information are used, while for image-only datasets, depth and camera information are estimated.

Other approaches, such as SpatialBot (Cai et al. 2025), provide both RGB and depth data from an image (in the form of RGB-D images) to the model and use the depth data in answering questions during inference. Depth-trained models achieve over 99% performance on the depth task in SpatialBench compared to approximately 84–90% for the RGB-only inputs, and about twice the end-to-end performance in pick-and-place experiment (Cai et al. 2025). However, this is not always the case. Using the same depth image to provide to GPT-4o in zero-shot results in a drop from 70.6 to 61.8 in the accuracy of the object’s position. Thus, the benefits of depth information depend on if the model is trained to utilize that information. The distinction also has implications for deployment cost. Running a model with depth added during dataset construction, adds no depth processing cost at inference time, while the model-time depth requires the presence of a depth sensor to gather information or an estimator and then processing it whenever the model is running. The answer to whether the runtime input for depth sensors is worth the cost appears to be somewhat dependent on the domain in which the models are deployed. Two datasets that have compared 2D versus 3D models indicate different results. RoboSpatial (Song et al. 2025) created a dataset that is ready for both 2D and 3D models and trains both types of models, showing a benefit to training with 3D models for robotics applications. Open3D-VQA (Zhang et al. 2025b) also created a 3D dataset but for aerial applications, where their 3D-LLM had better scores for estimating distance. But this distance advantage was not consistent across the other tested 3D models over 2D ones. SSR (Liu et al. 2025b) takes

a middle path of estimating depth from a monocular RGB image and compressing that depth into latent tokens for the model, allowing the model to benefit from depth information without requiring a depth sensor at inference time.

Video and Trajectories.

While static images can express spatial relations like “above” or “left of”, modeling dynamic spatial relations like “moving into” requires some notion of motion within the ground truth. Thus, there is a second main family of spatial datasets that was built with the goal of incorporating video supervision and trajectories. Such videos can enable new capabilities beyond static images alone, such as understanding the appearance of objects in a sequence of images (InternSpatial (Deng et al. 2026)), computing displacement of objects or tracking motion in videos (Video-STR (Wang et al. 2025)), and tracking objects that appear in various viewpoints (VILASR (Wu et al. 2025)). There are, actually, two different senses of the concept of “dynamic” relations. One sense is that the model’s input is a video itself (SpaceR (Ouyang et al. 2025), Video-STR (Wang et al. 2025), Embodied-R (Zhao et al. 2025)). The other is that the model’s reasoning is to be dynamic to the static image - for instance, the model may be provided with a static image but instead be asked what the scene will look like after it has been rotated (Spatial-DISE (Huang et al. 2026), Open3D-VQA (Zhang et al. 2025b)). Thus, these two different types of dynamic relation enable the model to provide answers about the motion it observed and the hypothetical transformations.

4.4 Source Environment

The environments that are represented in these spatial reasoning datasets closely reflect the target uses. Datasets intended for robotics prioritize indoor and tabletop environments, reflecting their intended tasks (Song et al. 2025; Chen et al. 2025a). RoboPoint also targets indoor environments, but does so almost exclusively using procedurally generated simulated environments, rather than real-world captures (Yuan et al. 2024). SpatialMosaic (Lee et al. 2025) covers both types of environments. It uses indoor scans from ScanNet++ and outdoor driving sequences from the Waymo Open Dataset specifically in order to avoid over-specialization to indoor environments. The dataset also provides both indoor and outdoor evaluations separately. General-purpose spatial reasoning datasets, on the other hand, sample from a wider variety of everyday environments, including indoor and outdoor scenes with people, objects, and activities (Deng et al. 2026; Cheng et al. 2024). Recent work expands on these environments by SpaceVista, featuring indoor, outdoor, and aerial (drone-view) video domains from a variety of scales (mm to km) in order to enable spatial reasoning across many physical scales (Sun et al. 2026).

5 Evaluation

After training an MLLM on spatial reasoning, it is necessary to assess whether the targeted spatial reasoning skill improved and that the general vision-language abilities of the model have not been harmed during the training process. While most surveyed works in the field report on both aspects of the evaluation, a notable minority of studies

do not report on the retention of general vision-language capability in addition to the improvement in the targeted capability. Table 2 summarizes the existing evaluation techniques for fine-tuning foundation models.

What is measured.

The evaluation of the VLMs targeted at spatial reasoning skills uses a variety of measures. While benchmarks report one overall score for spatial accuracy, they also provide separate scores for different categories of spatial reasoning skills such as perception, relations, and metric reasoning. SPAR-Bench, used alongside the SPAR-7M training dataset, outputs scores for depth and distance estimation, position matching, view-change inference, object relations, and spatial imagination (Zhang et al. 2025a). Spatial-DISE (Huang et al. 2026) separates spatial tasks into two cognitive axes: intrinsic (the internal parts and structure of a single object) vs. extrinsic (relations among several objects or to a frame) and cross-evaluates them with static vs. dynamic relations. Spatial-DISE finds that their models fall below the human baseline in different categories, including intrinsic-dynamic spatial relations, such as mentally rotating or folding one object. These failures correspond to the specific types of spatial estimations encoded in their datasets and could be averaged away by reporting a single number. Furthermore, the fine-tuned models are also evaluated on general vision-language benchmarks: SpaRE (Ogezi and Shi 2025) on MMMU, MMBench, and HallusionBench tests, and fDPO (Shen et al. 2025) on MME and POPE tests. While post-training helps improve performance on spatial tasks, performance on general tasks is not consistently retained in the reported results. SPAR’s score on MME decreases from 2186 to 2163 over its matched control and from 2323 to 2163 over the untrained base (Zhang et al. 2025a). For RoboSpatial, the MME-cognition score drops from 305.4 to 293.2, but all other general scores remain the same or increase (Song et al. 2025). Additionally, the trend of decreased performance across studies is difficult to follow due to variations in baselines. For example, fDPO’s general benchmark is evaluated relative to SpatialRGPT instead of its initial model and therefore does not indicate retention in its general capability (Shen et al. 2025). Spatial-DISE (Huang et al. 2026) has also found that fine-tuning on only portions of spatial data exhibits benefits only on external tasks that are of the same type (quadrant-specific experiments). For instance, if a model is only fine-tuned on viewpoint transformation then transfer to other less related tasks is “largely minimal or even detrimental”. However, when the model is fine-tuned on the entire of the DISE dataset then performance is improved on all external tasks. These findings imply that transfer is dependent on both the type of training performed and the type of evaluation tasks, and not that task similarity should be the sole explanation for transfer. General-benchmark results can only indicate retention relative to an appropriate initial checkpoint using the same evaluation protocol. While comparisons to other models can provide information about performance, they do not indicate what was gained or lost by the adapted model.

How it is scored.

The choice of metric is contingent upon the type of answer that the dataset asks. For datasets that require categorical answers (left/right, yes/no, or multiple choice),

accuracy and exact match are used to score the models (Zhang et al. 2025a; Huang et al. 2026). For answers that are provided as a metric (distance traveled, area of interest, etc.), exact match is often not a meaningful evaluation metric, so methods that consider whether the models’ answers fall within a certain tolerance of the ground truth values are used for evaluation; these can be metrics like $\pm 25\%$ relative-tolerance success rates (Cheng et al. 2024), or mean relative accuracy (Ouyang et al. 2025; Wu et al. 2025). For datasets that ask for open-ended answers, the answers from the models are provided to a language model to determine whether they are correct (Liu et al. 2025b; Deng et al. 2026). LLMs can also assist in judgment or numerical extraction alongside tolerance-based scoring (Cheng et al. 2024; Shen et al. 2025; Zhang et al. 2025b). For answers that ask for actions to be performed by a robot (points to be selected, trajectories to follow, etc.), there are two general approaches: some datasets use an algorithm to determine whether the predicted action is geometrically correct. For example, in RoboSpatial, a predicted 3D point is considered correct if it lies within the convex hull of a set of reference points computed from the scene geometry; the authors describe this as deliberately conservative (Song et al. 2025). The second approach is to evaluate whether the model succeeded with a downstream task of interest performed by the robot; i.e., RoboPoint (Yuan et al. 2024) and SpatialBot (Cai et al. 2025) ask whether the robot can complete a specific physical action after receiving an answer from the model. Finally, many of these datasets, such as Robo2VLM (Chen et al. 2025a) and Spatial-DISE (Huang et al. 2026), also provide human baselines to compare against human competence, not only against each other.

How generalization is tested.

Many recent evaluations have sought to separate in-domain from out-of-domain evaluations to determine whether the models are learning a generalizable skill or merely adapting to the structure of one particular benchmark, since a model can score well by memorizing benchmark-specific patterns like answer formats or phrasing without actually learning to reason spatially, and out-of-domain tests can reduce reliance on those familiar cues, making it harder to succeed without genuine generalization. SpatialLadder (Li et al. 2026) reports gains on benchmarks beyond their own splits, including CV-Bench, SPAR-Bench, and ViewSpatial-Bench. While these benchmarks vary in distribution and format, improvements across them do indicate some level of cross-benchmark transfer. However, they are not sufficient to completely eliminate the possibility of shared annotation practices, biases across tasks, or common sources of data. VILASR (Wu et al. 2025) evaluates on a variety of different task categories, spanning different input types and spatial reasoning demands, to demonstrate that the model is not reliant upon a particular regime of inputs or spatial condition. Furthermore, many of these models have evaluated their approach on real-world scenes to assess whether they can generalize their knowledge from being trained on synthetic scenes to the real world (Zhang et al. 2025b; Yuan et al. 2024). Finally, additional evaluations beyond accuracy can help to gauge generalizability. These range from lower-level checks, like testing sensitivity to wording by evaluating on paraphrased questions (Cheng et al. 2024), to broader ones like checking whether learning transfers to held-out environments (Song et al. 2025), or not using training videos

that overlap with the evaluation set to avoid measuring transfer across already-seen environments (Ouyang et al. 2025).

Table 2 Evaluation practices across the training datasets of Table 1. Codes are defined in the legend below.

Citation	Answer type / scoring	Retent.	Independ.	Robust.	Human
Song et al. (2025)	cat, act / acc, geom	✓	✓	env	–
Chen et al. (2025a)	cat / acc	–	–	–	✓
Ning et al. (2025)	cat, seg / acc, iou	–	–	–	–
Liu et al. (2025b)	cat, met, open / acc, judge	✓	–	–	–
Deng et al. (2026)	cat, met, open / acc, judge	✓	–	–	✓
Ranasinghe et al. (2024)	cat, act / acc, geom	✓	✓	–	–
Zhang et al. (2025b)	cat, met, open / acc, tol	–	–	s2r	–
Yuan et al. (2024)	act / geom, task	✓	✓	s2r	–
Ogezi and Shi (2025)	cat / acc	✓	✓	–	–
Tang et al. (2025)	cat, met / acc	✓	–	–	–
Huang et al. (2026)	cat / acc	–	–	–	✓
Ouyang et al. (2025)	cat, met / acc, mra	✓	✓	mod	–
Batra et al. (2025)	cat / acc	✓	–	–	–
Shen et al. (2025)	cat, met / acc, judge, tol	–	–	–	–
Wu et al. (2025)	cat, met / acc, mra	–	✓	–	–
Sarch et al. (2025)	cat / acc	✓	✓	–	–
Chen et al. (2025b)	cat / acc	✓	✓	–	–
Zhao et al. (2025)	cat / acc	✓	–	–	–
Chen et al. (2026)	cat, met / acc, mra	✓	–	–	–
Cai et al. (2025)	cat, met, act / acc, task	✓	–	–	–
Li et al. (2026)	cat, met / acc, mra	✓	–	–	–
Cheng et al. (2024)	cat, met / acc, tol	✓	–	para	–
Chen et al. (2024)	cat, met / acc, tol	✓	–	–	–
Wang et al. (2025)	cat, met / acc, mra	✓	✓	–	–
Sun et al. (2026)	cat, met / acc, mra	–	–	mod	✓
Zhang et al. (2025a)	cat, met / acc, mra	✓	–	–	✓
Lee et al. (2025)	cat, met / acc, tol	–	–	env	✓
Cai et al. (2026)	cat, met, open / acc, judge	✓	✓	dbias	–
Count		20/28	10/28	8/28	6/28

Answer type. Categorical (*cat*), metric (*met*), open-ended (*open*), action or point (*act*), segmentation (*seg*). A dash (–) denotes not reported, here and in every column.

Scoring. Accuracy or exact match (*acc*), relative tolerance band such as $\pm 25\%$ of the ground-truth value (*tol*), mean relative accuracy (*mra*), LLM-as-judge for semantic correctness (not only to extract an answer) (*judge*), geometric criterion such as convex-hull containment (*geom*), intersection-over-union (*iou*), downstream task success (*task*).

Retent. General-capability retention reported on non-spatial vision–language benchmarks with matched control. A checkmark denotes an assessment, not the absence of decline in performance.

Independ. The primary evaluation set is independent of the authors’ own training-data construction pipeline. Several entries also serve as benchmarks, for which independence is assessed against the accompanying training split.

Robust. Probe beyond headline accuracy: paraphrased questions (*para*), held-out environments (*env*), modality shift (*mod*), sim-to-real transfer (*s2r*), debiasing or shortcut probes such as no-vision and circular testing (*dbias*).

Human. Human task performance baseline reported. This is distinct from human-annotated ground truth, data verification, rating model outputs, and baselines from existing benchmarks.

Count. Final row; counts are over the 28 works listed.

6 Insights and Future Directions

6.1 Cross-Dataset Observations

When the datasets are considered collectively rather than separately, there are several patterns that do not emerge so readily from individual studies. For example, in some training phases, perceptual grounding of objects in an image appears to precede spatial reasoning about those same objects. This is not just a matter of style, but also yields better results in certain experiments. Studies such as SpatialLadder (Li et al. 2026) and Chen et al. (2025b) explicitly incorporate perceptual tasks (spatial awareness) before their reasoning-oriented training phases (such as spatial cognition). SpatialThinker (Batra et al. 2025) and Sparkle (Tang et al. 2025) provide evidence for the advantages of having foundational spatial awareness and rewards before more advanced training, although they do explore varying training orders. The convergence of these studies across different methods and systems implies that there is, in fact, value in having foundational spatial awareness, but they do not prove that a strict training order is required.

Relatedly, staging of the training of these models can lead to improved results at the corpus sizes of most of these datasets. Essentially, staging controls the order in which skills are introduced and built upon; for instance, training can be staged to introduce perceptual tasks first, supervised spatial reasoning second, and reinforcement learning third, so that each stage builds on the skills established by the one before it. This type of staging is implemented within the two-stage variant of Chen et al. (2025b), which trains perception before reasoning within reinforcement learning itself. SpatialLadder (Li et al. 2026), also contains three stages and moves from object localization to multi-dimensional spatial understanding to reinforcement learning with verifiable rewards. They report 43.9 on VSI-Bench for perception-then-spatial training, compared with 42.7 for spatial-only training and 40.7 for mixed training.

Another trend involves the balance between scale and supervision design. There have been successes with large corpora of synthetic data with broad coverage (Chen et al. 2024; Deng et al. 2026); however, SpatialThinker demonstrates that a significantly smaller corpus can also be effective when combined with verifiable spatial rewards. SpatialThinker is trained on $\sim 7,000$ QA samples and performs better than both supervised fine-tuning and conventional reinforcement-learning methods on fourteen tasks (Batra et al. 2025). While this benchmarking does not imply that rewards can always replace the need for large training data (since models and datasets differ), it does imply that alongside pretrained knowledge, carefully designed verifiable rewards can lessen the reliance on extremely large static training corpora.

The impact of reinforcement learning also varies based on when it is applied in the training pipeline. SpaceR benchmarks reward-based training against supervised fine-tuning from the same initial point and observes that reward-based training has higher overall performance across all tasks tested (Ouyang et al. 2025). However, in ViGoRL, it is found that applying GRPO to the base model results in less spatially grounded reasoning compared to the warm-start method (Sarch et al. 2025). Thus, reinforcement learning should not be thought of as uniformly better or substituting

for supervised learning. Instead, comparisons should only be made while holding constant the previous level of supervision and the point in the training pipeline at which reinforcement learning is applied.

As discussed in Section 4.3, depth and geometry are valuable as a supervision signal, while their value as runtime inputs relies on the task and the model. In datasets with large amounts of data, depth is often used during data collection and then training occurs using only RGB (Chen et al. 2024; Deng et al. 2026), which provides effective supervision. At inference time, providing geometry to the model is less clear, and the available evidence across domains is mixed: In robotics, geometry seems to provide benefits (Song et al. 2025), while in aerial images, Open3D-VQA (Zhang et al. 2025b) reports a distance-estimation advantage for 3D-LLM, while fine-tuned Qwen2-VL-7B has the higher overall score.

Finally, reference frames and viewpoints appear to be a common source of challenge, and their related terminology differs across datasets. Several datasets that test multiple frames or viewpoints consistently report that models struggle with frame transformations and perspective-taking, including RoboSpatial (Song et al. 2025) across ego, world, and object frames, Open3D-VQA (Zhang et al. 2025b) on allocentric-egocentric transformation, and InternSpatial (Deng et al. 2026) on multi-view reasoning. At the same time, terms such as “multi-view”, “egocentric”, and “object-centric” are used inconsistently across papers.

6.2 Evaluation Open Problems

While there has been a lot of work done, there still lacks a unified way to evaluate spatial reasoning and having shared evaluation metrics across works in Table 2. According to our observations, there are four issues that make it difficult to compare across works and reach conclusions about the efficacy of specific methods. Firstly, some works utilize LLMs to evaluate open-ended spatial answers (Liu et al. 2025b; Deng et al. 2026). However, the LLMs used might make mistakes in spatial interpretation or answer understanding. Additionally, agreement between the LLM evaluations and human spatial evaluations is not reported in these works. SSR compares a second judge, GPT-4o-mini, with its original evaluator. But this comparison does not imply agreement with human evaluations (Liu et al. 2025b).

Secondly, the threshold for what constitutes a “tolerated” error in the evaluation tasks is often defined separately in each paper. SpatialRGPT uses a relative error tolerance of $\pm 25\%$ of the ground-truth value (Cheng et al. 2024), and VILASR uses mean relative accuracy, which computes accuracy across ten relative accuracy thresholds from 0.50 to 0.95 (Wu et al. 2025). Because threshold values vary across papers, a model can exhibit higher accuracy simply by using a looser threshold, which makes accuracy values across different benchmarks not directly comparable. Thirdly, for some spatial benchmarks both the training and test sets are created using the same construction process. Because of this, improvements in accuracy observed between two benchmarks created in this way could be partly due to similarities in the construction process. Lastly, transfer changes with the training configuration and the spatial tasks used for evaluation. This means that high performance on benchmarks could indicate

only narrow performance on specific tasks rather than general spatial reasoning ability. Overall, evaluating accuracy on a single benchmark provides an incomplete view of a model’s spatial reasoning ability.

6.3 Future Directions

Building on these observations, we outline several directions for future work:

Unified dynamic annotation.

In Section 4.3 we discussed how motion datasets are split between those that require dynamic input (video) and those that only require dynamic reasoning across static images. Furthermore, motion, trajectory, visibility and affordance annotations are generally done via different and task-specific annotation systems (Wang et al. 2025; Yuan et al. 2024; Lee et al. 2025; Zhao et al. 2025). There is a need for a more unified system of dynamic annotation that can simultaneously capture spatial, temporal, and affordance information, as well as the physics that link these to space.

Efficient reasoning.

Reinforcement learning can improve multi-step spatial reasoning, but SpaceR shows that it can also produce long reasoning chains (Ouyang et al. 2025). These chains are expensive to decode, and high-resolution images also increase visual-token costs (Yang et al. 2025). Future datasets could include reasoning chains of different lengths, allowing models to reason more only when needed.

Toward more comparable evaluation.

In Section 5 existing evaluation methods are quite fragmented. Different works use different scoring methods, have different tolerance levels, use different benchmark datasets, and use LLM evaluators whose agreement with human spatial judgments is not reported. The improvements that are observed sometimes only transfer to similar tasks. A unified evaluation scheme using similar tasks, agreed tolerance ranges, common baselines, and clearly reported scoring procedures would enable more direct comparisons across datasets and would also help guide future research.

Spatial scale and geospatial data.

Most of the datasets discussed in this survey operate at what Montello (Montello 1993) refers to as figural and vista scale; that is, they contain information about small objects, tabletops, and rooms. While there are a few exceptions to this (such as SpaceVista’s ability to cover scales from mm to km (Sun et al. 2026), or Open3D-VQA’s aerial image understanding (Zhang et al. 2025b)), there are also a few separate datasets dedicated to geospatial tasks. For instance, there are datasets based upon remote sensing (Kuckreja et al. 2024; Li et al. 2024c) or geospatial chain-of-thought questions (Liu et al. 2026b), though these are generally separate from the types of datasets surveyed in this paper. How spatial reasoning can be transferred from one of these categories to the other, however, is an area for future research.

Human-centred settings.

Spatial understanding is important for safe human-robot interaction (Sisbot et al. 2007), yet the datasets we surveyed seldom capture its interactive dimension: a robot reasoning about space jointly with a person, where human intention and bidirectional interaction shape the spatial context. Datasets designed around such interaction would address a dimension that current embodied datasets largely leave open.

7 Conclusion

This survey examined spatial reasoning in fine-tuned MLLMs and VLMs from a data-centric point of view, with a special focus on various approaches used in the literature for annotation, design, and dataset building. We analyzed datasets used for post-training and classified the literature based on our taxonomy into three distinct categories: dataset design, building, and evaluation. Moreover, we organized them based on modality, annotation approach, framing, and view. Spatial reasoning ground-truth signals, perception and reasoning, and also framing in 3D space were comprehensively discussed from a data perspective. At the core of the paper, we carefully studied different dataset-building strategies and approaches, which give readers unique insights into vital design decisions before building datasets, given their applications and practical constraints in building large-scale datasets. After performing cross-dataset comparisons, we found several trends. Perception in some training curricula is established prior to higher-level spatial reasoning. There are also reports of benefit from staged training. Verifiable rewards may help reduce dependency on large fixed training sets in certain approaches. Depth and geometry are often helpful in creating supervision but show mixed results as runtime inputs. We conclude the survey by highlighting open problems and inconsistencies in evaluation protocols and possible future directions in spatial reasoning dataset development. These findings suggest that future progress depends not only on larger datasets, but also on more consistent task definitions, reference frames, supervision strategies, and evaluation protocols.

Statements and Declarations

Funding

We acknowledge the financial support of the Pan-Canadian AI Strategy, led by CIFAR, and the Alberta Machine Intelligence Institute (Amii).

Author contributions

Mahdieh Mallahnezhad: Conceptualization; Methodology; Investigation; Formal analysis; Writing—original draft; Writing—review & editing; Project administration. Masoud Jafaripour: Conceptualization; Investigation; Writing—original draft; Visualization; Writing—review & editing. Mohamadali Tofigh: Investigation; Formal analysis; Visualization; Writing—review & editing. Sheila Schoepp: Conceptualization; Investigation; Formal analysis; Visualization; Writing—review & editing. Osmar R. Zaiane: Conceptualization; Supervision; Writing—review & editing.

References

- Bai T, Liang H, Wan B, et al (2024) A survey of multimodal large language model from a data-centric perspective. CoRR abs/2405.16640
- Batra H, Tu H, Chen H, et al (2025) SpatialThinker: Reinforcing scene graph-grounded spatial reasoning via dense rewards. CoRR abs/2511.07403
- Cai W, Ponomarenko I, Yuan J, et al (2025) SpatialBot: Precise spatial understanding with vision language models. In: International Conference on Robotics and Automation (ICRA)
- Cai Z, Wang R, Gu C, et al (2026) Scaling spatial intelligence with multimodal foundation models. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)
- Chen B, Xu Z, Kirmani S, et al (2024) SpatialVLM: Endowing vision-language models with spatial reasoning capabilities. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)
- Chen K, Xie S, Ma Z, et al (2025a) Robo2VLM: Improving visual question answering using large-scale robot manipulation data. In: Conference on Neural Information Processing Systems (NeurIPS)
- Chen Y, Li L, Xi T, et al (2025b) Perception before reasoning: Two-stage reinforcement learning for visual reasoning in vision-language models. CoRR abs/2509.13031
- Chen Z, Zhang M, Yu X, et al (2026) Think with 3D: Geometric imagination grounded spatial reasoning from limited views. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)
- Cheng A, Yin H, Fu Y, et al (2024) SpatialRGPT: Grounded spatial reasoning in vision-language models. In: Conference on Neural Information Processing Systems (NeurIPS)
- Deng N, Gu L, Ye S, et al (2026) InternSpatial: A comprehensive dataset for spatial reasoning in vision-language models. In: International Conference on Learning Representations (ICLR)
- Feng J, Zeng J, Long Q, et al (2025) A survey of large language model-powered spatial intelligence across scales: Advances in embodied agents, smart cities, and earth science. CoRR abs/2504.09848
- Hegarty M (2004) Mechanical reasoning by mental simulation. Trends in Cognitive Sciences 8(6):280–285
- Huang X, He Q, Huang Z, et al (2026) Spatial-DISE: A unified benchmark for evaluating spatial reasoning in vision-language models. In: International Conference on

Learning Representations (ICLR)

- Jaton F (2021) Assessing biases, relaxing moralism: On ground-truthing practices in machine learning design and application. *Big Data & Society* 8(1):2053951721110135
- Johnson J, Hariharan B, van der Maaten L, et al (2017) CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*
- Kamath A, Hessel J, Chang K (2023) What’s “up” with vision-language models? investigating their struggle with spatial reasoning. In: *Conference on Empirical Methods in Natural Language Processing (EMNLP)*
- Kuckreja K, Danish MS, Naseer M, et al (2024) GeoChat: Grounded large vision-language model for remote sensing. In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*
- Lee J, Hockenmaier J (2025) Evaluating step-by-step reasoning traces: A survey. In: *Findings of the Association for Computational Linguistics, Conference on Empirical Methods in Natural Language Processing (EMNLP)*
- Lee K, Hong J, Lee S, et al (2025) Spatialmosaic: A multiview VLM dataset for partial visibility. *CoRR* abs/2512.23365
- Li F, Hogg DC, Cohn AG (2024a) Advancing spatial reasoning in large language models: An in-depth evaluation and enhancement using the stepgame benchmark. In: *AAAI Conference on Artificial Intelligence (AAAI)*
- Li F, Hogg DC, Cohn AG (2024b) Reframing spatial reasoning evaluation in language models: A real-world simulation benchmark for qualitative reasoning. In: *International Joint Conference on Artificial Intelligence (IJCAI)*
- Li H, Li D, Wang Z, et al (2026) SpatialLadder: Progressive training for spatial reasoning in vision-language models. In: *International Conference on Learning Representations (ICLR)*
- Li J, Li D, Savarese S, et al (2023) BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In: *International Conference on Machine Learning (ICML)*
- Li X, Ding J, Elhoseiny M (2024c) VRSBench: A versatile vision-language benchmark dataset for remote sensing image understanding. In: *Annual Conference on Neural Information Processing Systems (NeurIPS)*
- Liu D, Liang T, Hu Z, et al (2026a) Spatial intelligence in vision-language models: A comprehensive survey. *Artificial Intelligence Review*

- Liu F, Lin K, Li L, et al (2024) Mitigating hallucination in large multi-modal models via robust instruction tuning. In: International Conference on Learning Representations (ICLR)
- Liu H, Li C, Wu Q, et al (2023) Visual instruction tuning. In: Conference on Neural Information Processing Systems (NeurIPS)
- Liu J, Sun L, Fu R, et al (2026b) Towards faithful reasoning in remote sensing: A perceptually-grounded geospatial chain-of-thought for vision-language models. In: International Conference on Learning Representations (ICLR)
- Liu W, Xue Q, Wang H, et al (2025a) Spatial reasoning in multimodal large language models: A survey of tasks, benchmarks and methods. CoRR abs/2511.15722
- Liu Y, Ma M, Yu X, et al (2025b) SSR: Enhancing depth perception in vision-language models via rationale-guided spatial reasoning. In: Conference on Neural Information Processing Systems (NeurIPS)
- Ma X, Smart B, Bhalgat Y, et al (2024) When LLMs step into the 3D world: A survey and meta-analysis of 3D tasks via multi-modal large language models. CoRR abs/2405.10255
- Montello DR (1993) Scale and multiple psychologies of space. In: Conference on Spatial Information Theory (COSIT)
- Ning Z, Tian Z, Shi S, et al (2025) Enhancing spatial reasoning in multimodal large language models through reasoning-based segmentation. In: International Conference on Computer Vision (ICCV)
- Ogezi M, Shi F (2025) SpaRE: Enhancing spatial reasoning in vision-language models with synthetic data. In: Annual Meeting of the Association for Computational Linguistics (ACL)
- O’Keefe J, Nadel L (1979) Précis of O’Keefe & Nadel’s The hippocampus as a cognitive map. Behavioral and Brain Sciences 2(4):487–494
- Ouyang K, Liu Y, Wu H, et al (2025) SpaceR: Reinforcing MLLMs in video spatial reasoning. CoRR abs/2504.01805
- Piaget J, Inhelder B (2013) The Child’s Conception of Space. Routledge
- Ranasinghe K, Shukla SN, Poursaeed O, et al (2024) Learning to localize objects improves spatial reasoning in visual-LLMs. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)
- Sarch G, Saha S, Khandelwal N, et al (2025) Grounded reinforcement learning for visual reasoning. In: Conference on Neural Information Processing Systems (NeurIPS)

- Shen Y, Liu Y, Zhu J, et al (2025) Fine-grained preference optimization improves spatial reasoning in VLMs. In: Conference on Neural Information Processing Systems (NeurIPS)
- Sisbot EA, Marin-Urias LF, Alami R (2007) Spatial reasoning for human robot interaction. In: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
- Song CH, Blukis V, Tremblay J, et al (2025) RoboSpatial: Teaching spatial understanding to 2D and 3D vision-language models for robotics. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)
- Sun P, Lang S, Wu D, et al (2026) SpaceVista: All-scale visual spatial reasoning from mm to km. In: International Conference on Machine Learning (ICML)
- Tang Y, Qu A, Wang Z, et al (2025) Sparkle: Mastering basic spatial capabilities in vision language models elicits generalization to spatial reasoning. In: Findings of the Association for Computational Linguistics, Conference on Empirical Methods in Natural Language Processing (EMNLP)
- Tversky B (2005) Functional Significance of Visuospatial Representations, Cambridge University Press, pp 1–34. Cambridge Handbooks in Psychology
- Vieu L (1997) Spatial Representation and Reasoning in Artificial Intelligence, Springer, pp 5–41
- Wang W, Zou H, Luo T, et al (2025) Video-STR: Reinforcing MLLMs in video spatio-temporal reasoning with relation graph. CoRR abs/2510.10976
- Wu J, Guan J, Feng K, et al (2025) Reinforcing spatial reasoning in vision-language models with interwoven thinking and visual drawing. In: Conference on Neural Information Processing Systems (NeurIPS)
- Yang S, Li J, Lai X, et al (2025) VisionThink: Smart and efficient vision language model via reinforcement learning. In: Conference on Neural Information Processing Systems (NeurIPS)
- Yin S, Fu C, Zhao S, et al (2024) A survey on multimodal large language models. National Science Review 11(12):nwae403
- Yu S, Chen Y, Ju H, et al (2025) How far are VLMs from visual spatial intelligence? a benchmark-driven perspective. CoRR abs/2509.18905
- Yuan W, Duan J, Blukis V, et al (2024) RoboPoint: A vision-language model for spatial affordance prediction in robotics. In: Conference on Robot Learning (CoRL)
- Zha J, Fan Y, Yang X, et al (2025) How to enable LLM with 3d capacity? A survey of spatial reasoning in LLM. In: International Joint Conference on Artificial

Intelligence (IJCAI)

Zhang J, Chen Y, Zhou Y, et al (2025a) From flatland to space: Teaching vision-language models to perceive and reason in 3D. In: Datasets and Benchmarks Track, Neural Information Processing Systems (NeurIPS)

Zhang W, Zhou Z, Zeng X, et al (2025b) Open3D-VQA: A benchmark for embodied spatial concept reasoning with multimodal large language model in open space. In: ACM International Conference on Multimedia (ACM MM)

Zhao B, Wang Z, Fang J, et al (2025) Embodied-R: Collaborative framework for activating embodied spatial reasoning in foundation models via reinforcement learning. In: ACM International Conference on Multimedia (ACM MM)

Zheng X, Dongfang Z, Jiang L, et al (2025) Multimodal spatial reasoning in the large model era: A survey and benchmarks. CoRR abs/2510.25760