GeoMind: Explicit Spatial Reasoning via Dual-Reference Geometric Modeling

Xing Wei[1] , **Aoxiang Tian**[1] , **Shaofan Liu**[1*] , **Jiansheng Peng**[2,3] , **Chong Zhao**[1] , **Xiang Bi**[1] , **Yang Lu**[1] , **Benhong Zhang**[1] and **Fan Yang**[1]

[1]Hefei University of Technology, Hefei, 230009, China

[2]Department of Artificial Intelligence and Manufacturing, Hechi University, Hechi, 547000, China

[3]Key Laboratory of AI and Information Processing (Hechi University), Education Department of Guangxi Zhuang Autonomous Region

shaofanliu@hfut.edu.cn

Abstract

While Vision-Language Models (VLMs) excel at semantic understanding, they struggle to comprehend 3D spatial relationships from limited views. Their reliance on implicit geometric encoding often leads to severe hallucinations and inconsistencies in spatial reasoning tasks. To address this, we introduce GEOMIND, a *model-then-reason* framework that employs a single LLM to autoregressively generate an explicit **Geometric Description Language (GDL)** map, serving as a grounded context to derive the final answer. This intermediate GDL map provides an explicit and queryable world representation. Leveraging this explicit representation, we enforce a strict referential constraint, compelling the model to ground reasoning solely on the instantiated entities to ensure referential integrity and auditability. Specifically, we lift multi-view observations into object-centric tokens using frozen geometric priors and instance masks. The LLM is trained via a two-stage curriculum with programmatic supervision to generate the GDL map as a prerequisite for answering. On five spatial understanding benchmarks in both image and video settings, GEOMIND delivers average accuracy gains of **+6.9%** (2B) and **+9.8%** (8B) over Qwen3-VL baselines. Our results suggest that explicit geometric grounding enables robust spatial reasoning without human annotation, providing a scalable and practical route to stronger spatial intelligence in large VLMs.

1 Introduction

Spatial intelligence in embodied agents relies on constructing a coherent internal representation to support *mental completion* and *perspective taking* [Feng *et al.*, 2025; Wang *et al.*, 2023]. In recent years, Vision-Language Models (VLMs) have become a primary foundation for embodied agents to interpret visual observations and perform language-guided spatial reasoning [Yang *et al.*, 2025a]. However, while contemporary VLMs demonstrate strong semantic understanding, their spatial understanding remains substantially below human performance [Li *et al.*, 2025c].This gap may be partly related to a common design choice in VLMs to encode geometric cues implicitly within unstructured latent features [Wu *et al.*, 2024; Cheng *et al.*, 2024], which can hinder the formation of a stable spatial state under viewpoint changes. This limitation becomes most apparent during *reference-frame switching*: models struggle to reconcile view-invariant spatial facts (e.g., "the TV is behind the sofa") with transient visibility and occlusion evidence, often yielding viewpoint-inconsistent predictions and hallucinations across viewpoints [Li *et al.*, 2025a; Guan *et al.*, 2024]. This issue arises from the absence of an explicit, queryable world state. Without such a representation, models struggle to preserve viewpoint consistency and to reason over both allocentric topology and egocentric observations.

To bridge this gap, we introduce GeoMind, a framework that reformulates spatial QA as a generative process grounded in an explicit, *language-native* world model [Nejatishahidin *et al.*, 2024]. As illustrated in Figure 1, many VLM-based approaches produce answers directly from implicit latent representations (Figure 1.C, top). In contrast, our method uses a single LLM to autoregressively generate Geometric Description Language (GDL) tokens as an explicit intermediate map (Figure 1.B), and then generates the answer conditioned on this map (Figure 1.C, bottom). GDL features a *dual-reference* structure to address reference-frame switching: an *allocentric branch* encodes viewpoint-invariant topology, while an *egocentric branch* captures view-dependent observational evidence. This design separates "what exists in the world" from "what is currently observed," providing a structured scaffold for cross-view reasoning.

GeoMind is supported by a scalable programmatic data engine, which constructs GDL pseudo-labels by distilling geometric signals from reconstructed 3D scenes, enabling large-scale geometric supervision without expensive human annotation [Raychaudhuri and Chang, 2025]. Inference is performed under a self-consistency mechanism that constrains the model to derive answers solely from entities and relations instantiated in the GDL map. Under this strict grounding, the reasoning process becomes transparent and auditable. As shown in Figure 1, error cases can often be localized to omissions or inconsistencies in the intermediate map, and counterfactual edits to the GDL representation provide a direct way

*Corresponding author

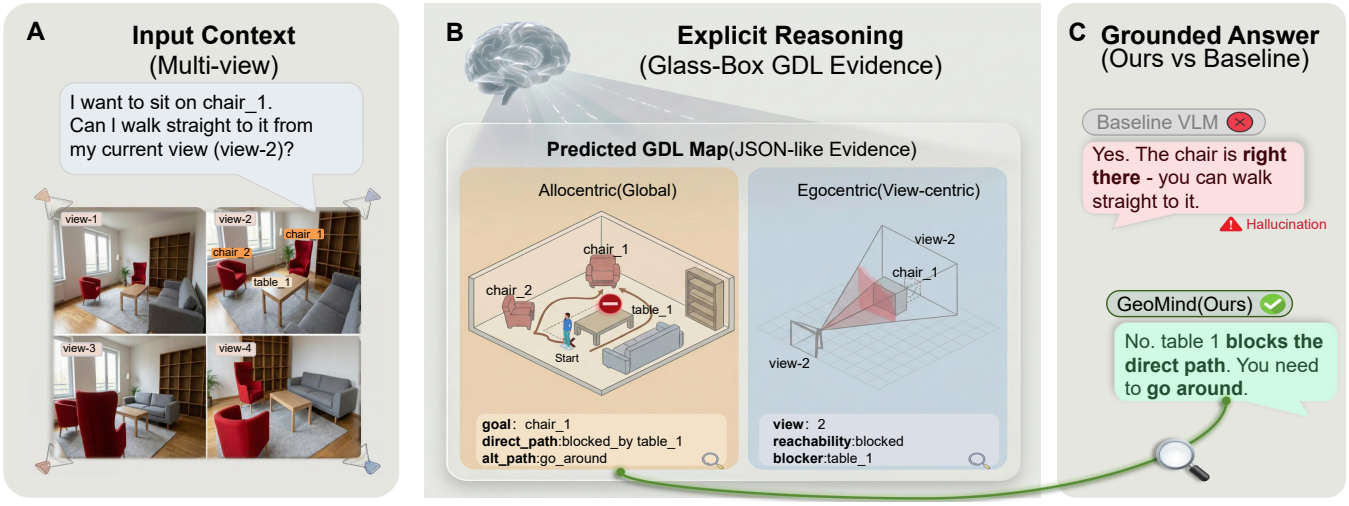

Figure 1: **The GeoMind Paradigm.** Unlike implicit end-to-end VLMs, we adopt a *model-then-reason* approach. Given multi-view inputs **(A)**, the model autoregressively generates an explicit Dual-Reference GDL map **(B)**, which disentangles allocentric topology from egocentric visibility, and then grounds the final answer **(C)** to improve self-consistency and auditability.

to test and revise the final answer.

Our main contributions are:

- **Dual-Reference GDL Framework.** We introduce a language-native cognitive map representation that explicitly disentangles allocentric (world-centric) topology from egocentric (view-dependent) visibility, providing a structured substrate for cross-view spatial reasoning.
- **Self-Consistent and Auditable Reasoning.** We enforce a self-consistent inference regime in which answers are derived solely from the generated GDL map, yielding coherent reasoning and enabling systematic error diagnosis through explicit map inspection and counterfactual editing.
- **Scalable Programmatic Supervision.** We propose an automated pipeline that distills 3D geometry into discrete symbolic supervision, enabling large-scale training without human annotation and yielding competitive performance against proprietary large-scale models on established spatial reasoning benchmarks.

2 Related Work

Current approaches to spatial reasoning span a spectrum from implicit adaptation to explicit modeling. The implicit stream optimizes the reasoning process itself, ranging from general VLMs (e.g., GPT-4V [Yang *et al.*, 2023], InternVL [Chen *et al.*, 2024]) to models fine-tuned with reinforcement learning (e.g., Spatial-SSRL [Liu *et al.*, 2025]). However, without explicit grounding in physical geometry, these methods often fail to maintain cross-view consistency in complex 3D scenarios. This has spurred a shift towards explicit modeling, where symbolic structures serve as reasoning scaffolds.

MindCube [Yin *et al.*, 2025] is a representative work in this direction, pioneering the "model-then-reason" paradigm by predicting 3D voxel states. Similar to other global-representation approaches like 3D-LLM [Hong *et al.*, 2023], it focuses primarily on *allocentric* (world-centric) topology. While effective for stable geometry, such global-only models often struggle with *situated reasoning*—tasks requiring a tight coupling between global position and local visibility (e.g., embodied navigation [Ma *et al.*, 2023]). GeoMind advances this paradigm by introducing Dual-Reference GDL, which explicitly disentangles two branches: an *allocentric branch* for stable topology, and an *egocentric branch* for local observational context (e.g., visibility, occlusion). This design provides a more rigorous scaffold for perspective-taking.

To construct such grounded representations, we bridge the gap between neural perception and symbolic mapping. While foundational vision models like VGGT [Wang *et al.*, 2025a] and DUSt3R [Wang *et al.*, 2024] excel at feed-forward reconstruction, they lack the semantic interface required for high-level reasoning. Conversely, traditional robotic mapping (SLAM) provides metric consistency but generates dense data not natively queryable by LLMs. Recent map-driven methods have explored converting BEV maps [Xu *et al.*, 2025] into language prompts, yet these are often tailored to domain-specific assumptions (e.g., autonomous driving). GeoMind synthesizes these threads: we leverage frozen geometric priors (VGGT) for robust 3D lifting and distill them into a lightweight map serialized as text tokens. This design allows our model to perform "in-context spatial reasoning" that combines the geometric rigor of SLAM with the semantic generalization of VLMs.

3 GeoMind

GeoMind is a neuro-symbolic *model-then-reason* framework for multi-view spatial QA. Instead of mapping pixels directly to answers, it introduces an explicit intermediate representation—the **Geometric Description Language (GDL)**—as a language-native cognitive map that encodes *allocentric* topology and *egocentric* evidence, enabling robust and auditable spatial reasoning.

Problem Formulation. Following standard multi-view spatial QA settings [Azuma *et al.*, 2022], given an observa-

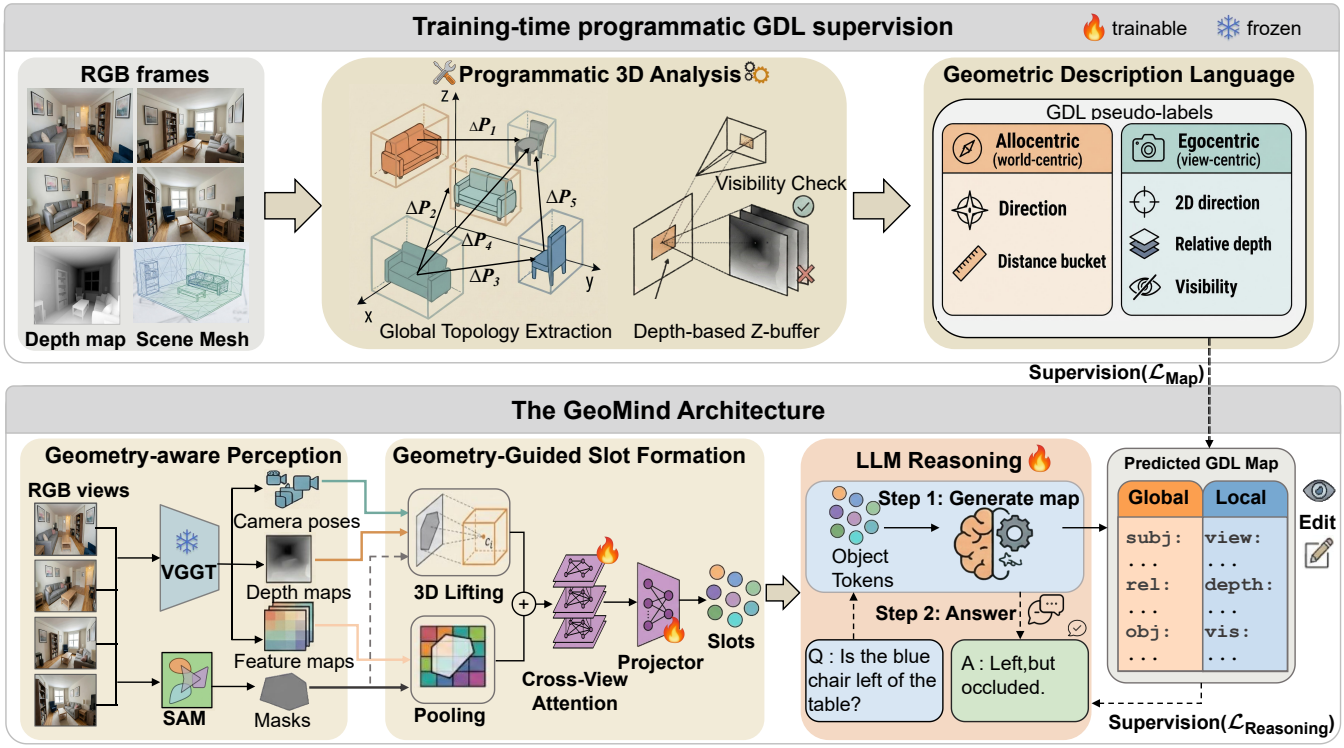

Figure 2: **The GeoMind Architecture.** (Top) Programmatic Supervision: 3D geometric ground truth is distilled into discrete GDL pseudo-labels to supervise ($\mathcal{L}_{\text{Map}}$) the allocentric and egocentric branches. (Bottom) Inference Pipeline: (1) Perception extracts visual features and geometric priors (Depth/Pose) via frozen VGGT and SAM; (2) Slot Formation fuses inputs through masked pooling, 3D lifting, and trainable Cross-View Attention to yield object tokens; (3) LLM Reasoning autoregressively generates the explicit GDL Map (Step 1) followed by the grounded answer (Step 2), enforcing spatial self-consistency.

tion sequence $\mathcal{V} = \{I_k\}_{k=1}^{K}$ and a query q, our goal is to predict the answer a. Unlike end-to-end approaches that model $P(a \mid \mathcal{V}, q)$ directly, we introduce an explicit intermediate map variable $\mathcal{M}$ (instantiated as GDL) to enforce geometric consistency. This factorizes the inference into map construction and map-based reasoning:

$$P(a \mid \mathcal{V}, q) \approx \sum_{\mathcal{M}} P(a \mid \mathcal{M}, q)\, P(\mathcal{M} \mid \mathcal{V}). \tag{1}$$

In practice, we decode a MAP estimate $\hat{\mathcal{M}} = \arg\max_{\mathcal{M}} P(\mathcal{M} \mid \mathcal{V})$ and generate a conditioned on $(\hat{\mathcal{M}}, q)$.

3.1 Geometry-Aware Object Abstraction

Our perception module uses a frozen VGGT encoder to extract geometry-aware patch features with cross-view consistency. Given K views $\mathcal{V} = \{I_k\}_{k=1}^{K}$, the encoder outputs patch features $\mathbf{F}_{\text{patch}} \in R^{K \times N_p \times C}$, together with per-view depth maps $\{\mathbf{Z}_k\}_{k=1}^{K}$ and relative camera poses $\{\hat{\mathbf{T}}_k\}_{k=1}^{K}$, where N_p denotes the number of patch tokens per view. During training, ground-truth depth and poses are used only to construct reliable 3D lifting and supervision signals. At inference, we perform 3D lifting using $\{\mathbf{Z}_k, \hat{\mathbf{T}}_k\}$ without requiring external pose inputs [Miao *et al.*, 2025].

To preserve spatial layout information during patch tokenization, we inject learnable positional encodings $\mathbf{E}_{\text{pos}}$ into the feature stream:

$$\mathbf{F}_{\text{in}} = \mathbf{F}_{\text{patch}} + \mathbf{E}_{\text{pos}}, \tag{2}$$

where $\mathbf{E}_{\text{pos}} \in R^{K \times N_p \times C}$ encodes each patch's 2D location within its image. The frozen VGGT backbone supplies cross-view coherence, while $\mathbf{E}_{\text{pos}}$ retains local spatial context for subsequent object abstraction.

Geometry-Guided Slot Formation We compress multi-view inputs into *object slots*. These object-level tokens represent merged 3D instances and serve as concise visual prompts for subsequent GDL generation ($N \ll K \times N_p$). We adopt a lift-and-merge pipeline that synergizes mask-based proposals with 3D multi-view consistency.

Specifically, for each view index $k \in \{1, \ldots, K\}$, we obtain a set of instance masks $\mathcal{S}_k = \{s_{k,j}\}_{j=1}^{J_k}$, where j indexes mask proposals in view k and J_k is the number of masks. We utilize ground-truth masks with stochastic augmentations (e.g., erosion, noise injection) during training to mitigate the domain gap, while switching to SAM for open-world inference [Ravi *et al.*, 2024].

For each mask $s_{k,j}$, we compute an appearance descriptor

$\mathbf{f}_{k,j} \in R^C$ via masked average pooling:

$$\mathbf{f}_{k,j} = \frac{1}{|s_{k,j}|} \sum_{(u,v)\in s_{k,j}} \mathbf{F}_{\text{in}}^{(k)}(u,v). \tag{3}$$

Simultaneously, we back-project masked pixels to the world frame using $\mathbf{Z}_k$ and $\hat{\mathbf{T}}_k$:

$$\mathbf{x}_{\text{world}} = \hat{\mathbf{T}}_k \mathbf{Z}_k(u,v)\mathbf{K}_k^{-1}[u,v,1]^\top, \tag{4}$$

where $\mathbf{K}_k$ is the camera intrinsic matrix. The collection of lifted points forms a candidate point cloud $\mathcal{P}_{k,j}$, from which we derive an axis-aligned 3D bounding box $\mathbf{B}_{k,j}$. Thus, each proposal is characterized by both a semantic descriptor $\mathbf{f}_{k,j}$ and a spatial support $\mathbf{B}_{k,j}$.

Cross-View Correspondence. To associate proposals across views, we construct a pairwise cost matrix that balances geometric overlap and appearance affinity. We normalize the cosine distance to $[0,1]$ and define the matching cost as:

$$\begin{aligned} c_{\text{geo}} &= 1 - \text{IoU}_{3D}(\mathbf{B}_{k,j}, \mathbf{B}_{k',j'}), \\ c_{\text{app}} &= \tfrac{1}{2}\big(1 - \cos(\mathbf{f}_{k,j}, \mathbf{f}_{k',j'})\big), \\ \mathcal{C}\big((k,j),(k',j')\big) &= \alpha\, c_{\text{geo}} + (1-\alpha)\, c_{\text{app}}. \end{aligned} \tag{5}$$

We solve for optimal one-to-one assignments using the Hungarian algorithm [Kuhn, 1955]. To enforce spatial feasibility, we apply a gating mechanism where pairs with $\text{IoU}_{3D} \le \tau_{\text{iou}}$ are assigned infinite cost. Matched proposals are then merged into persistent 3D instances $\{O_i\}$.

Slot Encoding and Projection. For each merged 3D instance O_i, let $\mathcal{I}_i$ denote the set of indices for matched mask proposals belonging to O_i. We aggregate their per-view descriptors $\{\mathbf{f}_{k,j}\}_{(k,j)\in\mathcal{I}_i}$ into an instance-level appearance embedding $\bar{\mathbf{f}}_i \in R^C$ via cross-view attention pooling:

$$\bar{\mathbf{f}}_i = \sum_{(k,j)\in\mathcal{I}_i} \alpha_{k,j}^{(i)}\, \mathbf{f}_{k,j}, \quad \alpha_{k,j}^{(i)} = \frac{\exp(s_{k,j}^{(i)})}{\sum_{(k',j')\in\mathcal{I}_i} \exp(s_{k',j'}^{(i)})}, \tag{6}$$

where $s_{k,j}^{(i)} = \mathbf{q}_{\text{att}}^\top \mathbf{W}_a\, \mathbf{f}_{k,j}$ is the attention score derived from a shared learnable query $\mathbf{q}_{\text{att}}$ and projection $\mathbf{W}_a$. We further explicitly attach 3D geometry, concatenating the centroid $\mathbf{c}_i \in R^3$ and box extent $\mathbf{s}_i \in R^3$. A lightweight projector ϕ (implemented as an MLP) maps this representation to the LLM input space:

$$\mathbf{z}_i = \phi\big([\bar{\mathbf{f}}_i; \mathbf{c}_i; \mathbf{s}_i]\big). \tag{7}$$

The resulting object tokens $\mathcal{Z} = \{\mathbf{z}_i\}_{i=1}^N$ are prepended to the LLM sequence as visual prompts for subsequent GDL generation.

3.2 GDL: A Dual-Frame Cognitive Map

We represent the latent map variable $\mathcal{M}$ as a Geometric Description Language (GDL) sequence. GDL serves as a compact, text-serializable scene abstraction that the LLM can read and manipulate directly. As illustrated in Figure 3, a GDL sample consists of (i) an *object registry* that assigns a unique identifier and a category label to each instance, and (ii) a set of pairwise relations written in two complementary frames.

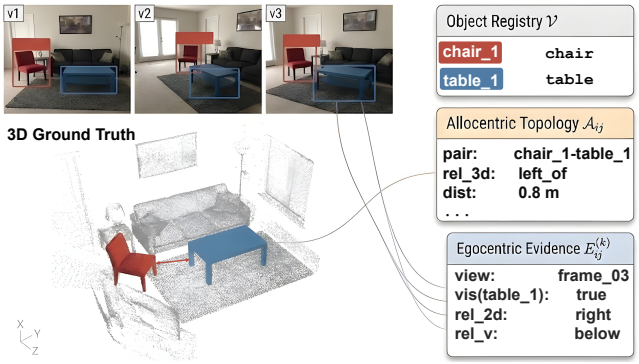

Figure 3: **Structure of the Geometric Description Language (GDL).** Left: Observed Scene. Multi-view RGB observations (top) and the associated 3D point cloud (bottom). Right: Symbolic Map. The serialized GDL, in which the Allocentric Branch (orange) encodes viewpoint-invariant topology (e.g., `left_of`), and the Egocentric Branch (blue) specifies frame-conditioned evidence (e.g., `vis: true`). Connecting lines indicate the explicit grounding between discrete tokens and physical entities.

The *allocentric* branch encodes viewpoint-invariant spatial topology in a shared world frame, such as coarse relative direction and distance (e.g., `left_of`). The *egocentric* branch records frame-conditioned evidence tied to a specific observation, such as visibility and view-dependent ordering (e.g., `vis: true`). Keeping both branches makes the representation explicit about what is stable across views and what is supported by a particular view, which is central to perspective taking and cross-view consistency.

Dual-Frame Relation Schema

We use a right-handed world frame with $+Y$ aligned with gravity and measure distances in meters; allocentric directions are defined on the XZ plane. GDL defines a relational graph over object nodes $\mathcal{O} = \{o_i\}_{i=1}^N$. Each node is associated with a unique identifier and a category token (e.g., `"chair"`), providing a stable reference for subsequent relations.

For each ordered pair (i,j), the relation is written as $R_{ij} = \langle A_{ij}, \{E_{ij}^{(k)}\}_k \rangle$, where A_{ij} denotes allocentric topology and $E_{ij}^{(k)}$ denotes egocentric evidence for view k. Figure 3 shows a concrete serialized example.

Allocentric Topology (world-centric). A_{ij} encodes viewpoint-invariant spatial memory in a unified world coordinate system. It captures stable 3D topology (e.g., relative direction, distance, and height ordering), forming a persistent map that remains consistent under viewpoint changes.

Egocentric Evidence (view-centric). $E_{ij}^{(k)}$ records view-dependent evidence of the same relation in a specific observation k. It captures local projection states such as visibility, relative 2D configuration, and ordinal depth, enabling perspective-taking and grounding the map against what is visually supported in each view.

The dual-frame structure maintains a globally consistent map and grounds it in view-specific evidence. Each merged 3D instance O_i is assigned a unique object identifier (e.g., `obj_i`) in the serialized GDL, and all relations and evidential records reference objects solely through these identifiers. During reasoning, the model is restricted to using the same identifiers via explicit in-context constraints and post-hoc validation, which prevents entity drift even when category names are ambiguous.

3.3 Programmatic Supervision Pipeline

To obtain supervision without manual annotation, we derive *geometry-based* GDL targets from the 3D reconstructions provided by ScanNet v2 [Dai *et al.*, 2017] on the training set.

Scene Parsing. Following standard 3D scene graph definitions [Wald *et al.*, 2020], we parse the semantic mesh into an object set $\mathcal{O} = \{o_i\}_{i=1}^{N_{\text{GT}}}$, where N_{GT} denotes the instance count. Each object o_i is parameterized by its 3D centroid $\mathbf{c}_i$, extent $\mathbf{s}_i$, semantic label y_i, and a set of geometry keypoints $\mathcal{Q}_i$.

Allocentric Relation Graph. To avoid the computational overhead of dense fully-connected graphs [Zhou *et al.*, 2019], we construct a sparse directed (i.e., ordered) relation graph $\mathcal{G} = (\mathcal{O}, \mathcal{E})$ over mesh instances. For each source object o_i, we define a neighbor set $\mathcal{N}(i)$ as its M_{nbr} nearest objects measured by Euclidean distance between centroids, and add edges (i, j) for all $j \in \mathcal{N}(i)$. This yields $|\mathcal{E}| = O(N_{\text{GT}} \cdot M_{\text{nbr}})$ edges, reducing the supervision size from quadratic to linear in N_{GT}. For each retained edge, we compute the displacement $\Delta\mathbf{p}_{ij} = \mathbf{c}_j - \mathbf{c}_i$ and discretize its azimuth on the XZ plane and metric distance into categorical bins to obtain allocentric relation labels (Sec. 3.2).

Egocentric Visibility Validation. To derive view-dependent evidence, we use ground-truth camera extrinsics $\mathbf{T}_k$ and intrinsics $\mathbf{K}_k$ to project 3D keypoints into each view[Ravi *et al.*, 2020]. A centroid-only baseline is brittle under partial occlusion and depth outliers (e.g., missing or noisy measurements around the projected centroid). We therefore adopt multi-keypoint validation, which improves robustness by requiring depth-consistent support for at least one keypoint. For each $\mathbf{p} \in \mathcal{Q}_i$, we project it to pixel $\mathbf{u}_p$ with projective depth z_p, and compare z_p with the sensor depth map $\mathbf{Z}_k$:

$$v_i^{(k)} = \frac{1}{|\mathcal{Q}_i|} \sum_{\mathbf{p} \in \mathcal{Q}_i} I(\mathbf{u}_p \in \Omega \wedge |z_p - \mathbf{Z}_k(\mathbf{u}_p)| < \tau_z), \quad (8)$$

where Ω denotes the valid image domain. We mark o_i as `visible` in view k if $v_i^{(k)} > 0$. Unless otherwise specified, we set $\tau_z = 3.0\,\text{m}$ to tolerate reconstruction misalignment and sensor noise. For visible pairs, we further compute 2D relations (e.g., `left`/`overlap`) from projected keypoint spans and infer ordinal depth (`nearer`/`farther`) by comparing median depths of validated keypoints.

Serialization. Finally, the allocentric relations and egocentric evidences are serialized into the GDL schema (Sec. 3.2), forming the ground-truth sequence $\mathcal{M}_{\text{GT}}$ for the generative objective $\mathcal{L}_{\text{Map}}$.

3.4 Two-Stage Training Protocol

We formulate both cognitive map construction and question answering as conditional sequence generation. Training follows a two-stage protocol that first stabilizes GDL generation, and then introduces reasoning supervision while preserving map fidelity.

Stage 1: Cognitive Map Construction. Given multi-view observations $\mathcal{V}$, we train the model to generate the ground-truth GDL map $\mathcal{M}_{\text{GT}} = (m_1^{\text{GT}}, \ldots, m_T^{\text{GT}})$ by maximizing its likelihood:

$$\mathcal{L}_{\text{Map}} = -\sum_{t=1}^{T} \log P\big(m_t^{\text{GT}} \mid m_{<t}^{\text{GT}}, \mathcal{V}\big). \quad (9)$$

This stage focuses on learning a valid and structurally consistent map representation from visual inputs.

Stage 2: Joint Map-and-Reasoning Training. We then train the model to answer the query q conditioned on the ground-truth map $\mathcal{M}_{\text{GT}}$ (teacher forcing):

$$\mathcal{L}_{\text{Reasoning}} = -\sum_{s=1}^{S} \log P\big(a_s^{\text{GT}} \mid a_{<s}^{\text{GT}}, \mathcal{M}_{\text{GT}}, q\big), \quad (10)$$

where $a_{\text{GT}} = (a_1^{\text{GT}}, \ldots, a_S^{\text{GT}})$ is the tokenized answer. The overall objective in Stage 2 is

$$\mathcal{L}_{\text{Joint}} = \mathcal{L}_{\text{Map}} + \lambda\, \mathcal{L}_{\text{Reasoning}}. \quad (11)$$

We use λ to balance map preservation and reasoning adaptation; implementation details are provided in Sec. 4.1.

At test time, the model performs sequential generation where it first predicts a map $\hat{\mathcal{M}}$ from $\mathcal{V}$ and subsequently derives the answer conditioned on $(\hat{\mathcal{M}}, q)$.

4 Experiments

4.1 Experimental Setup

Benchmark Overview. We evaluate GeoMind on five spatial reasoning benchmarks that cover both image and video settings. Image-based benchmarks include SITE [Wang *et al.*, 2025b], ViewSpatialBench [Li *et al.*, 2025a], MindCube [Yin *et al.*, 2025], and MMSI-Bench [Yang *et al.*, 2025d]. VSI-Bench [Yang *et al.*, 2025b] targets route planning in egocentric videos. For programmatic supervision, we use the ScanNet v2 [Dai *et al.*, 2017] training split (1,513 scenes) and generate physics-based GDL labels via mesh analysis. For reasoning alignment, we aggregate ScanQA [Azuma *et al.*, 2022], ScanRefer [Chen *et al.*, 2020], and Nr3D [Achlioptas *et al.*, 2020] and restrict all samples to the same training scenes to prevent leakage.

Evaluation Metrics. We follow the official evaluation protocols for each benchmark. For multiple-choice tasks in VSI-Bench, we report MRA, the Multiple-choice Response Accuracy. For SITE, we report CAA, the Correct Answer Accuracy, which checks both the answer text and the grounding bounding box. For open-ended QA benchmarks, we report exact match accuracy (Acc) under deterministic decoding. When required by the benchmark protocol, we filter semantic equivalents via LLM-based evaluation where applicable.

Model	Image-based Tasks				Video	Avg.
	SITE	ViewSpatial	MindCube	MMSI	VSI-Bench	
Metric	**CAA**	**Acc**	**Acc**	**Acc**	**MRA**	
Proprietary Models						
GPT5[Singh *et al.*, 2025]	61.8	45.5	56.3	41.8	55.0	52.1
Gemini-2.5-Pro[Team *et al.*, 2025]	57.0	46.0	57.6	38.0	53.5	50.4
Grok-4[xAI, 2025]	47.0	43.2	63.5	37.8	47.9	47.9
Seed-1.6[Guo *et al.*, 2025]	54.6	43.8	48.7	38.3	49.9	47.1
Open-source General VLMs						
Qwen2.5-VL-7B[Bai *et al.*, 2025]	37.6	36.8	36.0	26.8	32.3	33.9
Qwen2.5-VL-3B[Bai *et al.*, 2025]	33.1	31.9	37.6	28.6	27.0	31.6
InternVL3-8B[Zhu *et al.*, 2025]	41.1	38.6	41.5	28.0	42.1	38.3
InternVL3-2B[Zhu *et al.*, 2025]	30.0	32.5	37.5	26.5	32.9	31.9
Spatial-specialized Models						
Cambrian-S-7B[Yang *et al.*, 2025c]	33.0	40.9	39.6	25.8	67.5	41.4
Cambrian-S-3B[Yang *et al.*, 2025c]	28.3	39.0	32.5	25.2	57.3	36.5
ViLaSR-7B[Wu *et al.*, 2025]	38.7	35.7	35.1	30.2	44.6	36.9
SpatialLadder-3B[Li *et al.*, 2025b]	27.9	39.8	43.4	27.4	44.8	36.7
SpaceR-7B[Ouyang *et al.*, 2025]	34.2	35.8	37.9	27.4	41.5	35.4
MindCube-3B-RawOA-SFT[Yin *et al.*, 2025]	6.3	24.1	51.7	1.7	17.2	20.2
GeoMind Models						
Qwen3-VL-2B (baseline)[QwenLM Team, 2025]	35.6	36.9	34.5	28.9	50.3	37.2
GeoMind-2B (Ours)	**54.1**	**39.8**	**39.2**	**32.8**	**54.6**	**44.1**
Δ *vs. Qwen3-VL-2B*	*+18.5*	*+2.9*	*+4.7*	*+3.9*	*+4.3*	*+6.9*
Qwen3-VL-8B (baseline)[QwenLM Team, 2025]	45.8	42.2	29.4	31.1	57.9	41.3
GeoMind-8B (Ours)	**62.3**	**50.4**	**46.8**	**36.2**	**59.8**	**51.1**
Δ *vs. Qwen3-VL-8B*	*+16.5*	*+8.2*	*+17.4*	*+5.1*	*+1.9*	*+9.8*

Table 1: Performance comparison on five spatial reasoning benchmarks. Our GeoMind models (based on Qwen3-VL backbone) achieve state-of-the-art results among open-source models and are competitive with proprietary models. Metrics: CAA (Correct Answer Accuracy) for SITE, Acc (Accuracy) for others, MRA (Multiple-choice Response Accuracy) for VSI-Bench.

Implementation Details. We build GeoMind on the Qwen3-VL backbone with frozen vision and language components. Trainable modules include the object projector (3.9M parameters) and LoRA layers (rank 64, alpha 128, targeting attention projections, 40.4M parameters), totaling 44.3M parameters. Training employs AdamW with a two-stage curriculum over 15 epochs: the first 10 epochs focus on map generation (learning rates 5×10^{-6} for LoRA and 5×10^{-5} for the projector), followed by 5 epochs of joint training with $\mathcal{L}_{\text{Joint}} = \mathcal{L}_{\text{Map}} + 0.3\mathcal{L}_{\text{Reasoning}}$. We use an effective batch size of 8 with cosine annealing on 4 NVIDIA A100 (80GB) GPUs using BFloat16 precision, setting $N_{\text{slots}} = 64$ and $M_{\text{nbr}} = 5$ by default.

Model	ViewSpatial-Bench				VSI
	Camera		*Person*		
	R.Dir	O.Ori	O.Ori	R.Dir	Route
InternVL3-8B	50.3	27.5	42.5	37.5	42.1
GPT-5	60.1	27.9	40.9	48.4	55.0
Qwen3-8B	54.2	29.7	47.2	40.2	57.9
GeoMind-8B	**56.4**	**40.7**	**58.6**	**53.9**	**64.8**
Improv.	*+2.2*	*+11.0*	*+11.4*	*+13.7*	*+6.9*

Table 2: **Fine-grained Analysis.** Breakdown of **ViewSpatial** perspective-taking and **VSI** planning performance. GEOMIND minimizes the Camera-to-Person gap.

4.2 Main Results

Overall Performance. Table 1 summarizes the quantitative results. GEOMIND demonstrates superior performance among the compared open-weight baselines. Notably, **GeoMind-8B** achieves an average accuracy of **51.1%**, surpassing Gemini-2.5-Pro (50.4%) and narrowing the gap to GPT-5 (52.1%) to within **1.0 percentage point**. Most strikingly, on perspective-sensitive benchmarks like ViewSpatial and VSI-Bench, our 2B model substantially outperforms the spatial-specialized baseline MindCube-3B by a 37.4-point absolute margin.

Analysis of Perspective Taking. To understand the source of these gains, we break down performance on reference-frame switching tasks (Table 2). Standard VLMs struggle

Variant	Overall	ViewSpatial		VSI
		Cam.	Per.	Route
Full Model	**51.1**	**48.6**	**51.6**	**64.8**
Impact of GDL Structure				
(a) w/o GDL	44.7 (-6.4)	42.3	38.7	52.1
(b) w/o Allocentric	47.8 (-3.3)	47.1	41.5	51.2
(c) w/o Egocentric	48.5 (-2.6)	43.2	47.8	54.5
Impact of Training & Inference				
(d) w/o Two-Stage	47.9 (-3.2)	45.8	44.2	52.8
(e) w/o Self-Consistency	49.2 (-1.9)	47.2	46.8	54.1

Table 3: **Ablation Study on GeoMind-8B.** Accuracy (%) on **validation sets**. "Overall" averages all five benchmarks. *Route* refers to the route-planning subtask in VSI-Bench.

to generalize from camera-centric inputs to person-centric queries, exhibiting a sharp drop (e.g., Qwen3: 54.2% $\rightarrow$ 40.2%). In contrast, GeoMind maintains robustness, achieving **53.9%** on the challenging *Person-Rel.Dir* task. This confirms that our Allocentric Branch effectively disentangles global topology from local visibility.

Quality of Intermediate Representation. We verify that the intermediate GDL acts as a trustworthy world model. On the ScanNet validation set, GeoMind achieves a **79.8%** valid JSON format rate and **84.2%** structural consistency. Qualitative analysis further shows that the model correctly captures **87.3%** of salient spatial relations, enabling auditable reasoning.

4.3 Ablation and Analysis

To quantify the contribution of each component, we ablate GeoMind-8B on validation sets (Table 3). Direct QA without GDL yields the largest degradation (**-6.4%** Overall) and sharply reduces Route accuracy (64.8 $\rightarrow$ 52.1). This confirms that performance gains stem from reasoning over an explicit intermediate map rather than learning shortcut correlations. Removing the Allocentric Branch disproportionately hurts person-reference queries (51.6 $\rightarrow$ 41.5), while removing the Egocentric Branch mainly degrades camera-reference grounding (48.6 $\rightarrow$ 43.2). This outcome validates our dual-reference design: allocentric topology is essential for reference-frame switching, whereas egocentric evidence stabilizes view-conditioned grounding.

Training and Consistency. Skipping the two-stage training curriculum decreases Overall accuracy by 3.2 points, suggesting that aligning map structure prior to joint optimization improves stability. Disabling self-consistent decoding further reduces accuracy by 1.9 points, indicating that map-grounded generation effectively mitigates error propagation.

Hyperparameter Sensitivity. We further analyze model robustness in Table 4. Regarding Slot Capacity, performance peaks at $N_{\text{slots}} = 64$. Smaller budgets (e.g., $N = 30$) degrade performance significantly, likely due to the failure to capture all relevant entities in cluttered environments. Consequently, we establish 64 as a conservative upper bound to prevent truncation. Increasing N_{slots} further yields diminishing returns (51.1% vs. 50.9%), as it primarily improves coverage redundancy without enhancing the effective relation density.

Hyperparameter	Values (Acc %)					
Max Slots (N_{slots})	24 (48.7)	32 (49.2)	48 (50.5)	**64*** (**51.1**)	96 (51.0)	128 (50.9)
Neighbors (M_{nbr})	1 (48.8)	3 (49.5)	**5*** (**51.1**)	7 (51.0)	10 (50.8)	15 (50.4)

Table 4: **Hyperparameter sensitivity (GeoMind-8B).** We report the average accuracy (%) across five benchmarks. Default settings are marked with *.

In terms of Graph Sparsity, setting $M_{\text{nbr}} = 5$ proves optimal. While overly sparse graphs risk fragmenting the topology and hindering multi-hop propagation, excessive connectivity ($M_{\text{nbr}} \geq 10$) introduces irrelevant long-range noise that dilutes geometric distinctiveness. This validates the necessity of a balanced and sparse topological representation for precise spatial routing.

4.4 Visualization of Reasoning Process

We visualize the inference trajectory of GeoMind in Figure 4. The model first converts multi-view inputs into explicit GDL tokens, where the allocentric branch encodes stable spatial topology and the egocentric branch records view-dependent evidence such as visibility and occlusion. For queries requiring mental rotation, the reasoning module uses these grounded geometric relations to infer the correct orientation. The examples show that the symbolic scaffold supports auditable spatial reasoning beyond simple pattern matching.

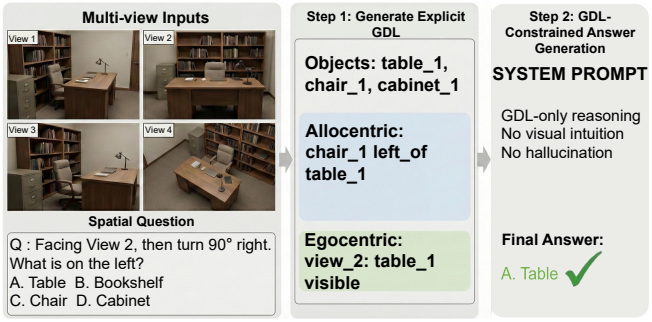

Figure 4: **Qualitative visualization.** GeoMind generates an explicit GDL map to strictly ground the reasoning process. This structured scaffold enables accurate perspective-taking and mental rotation (e.g., inferring relative orientation) while ensuring transparent auditability.

5 Conclusion

We introduced GeoMind to ground spatial reasoning in an explicit, dual-reference map. By constraining inference to self-generated entities, our approach mitigates geometric hallucinations and enables transparent auditability via map inspection. Empirically, GeoMind achieves performance competitive with proprietary models, confirming that explicitly reconciling allocentric topology with egocentric evidence is essential for robust perspective-taking.

Acknowledgments

This work was supported in part by the Fundamental Research Funds for the Central Universities of China (Grant No. PA2025GDSK0036), the National Natural Science Foundation of China (Grant No.62563009), the Anhui Provincial-Level Science and Technology Innovation Tackle Plan Projects (Grant Nos.202523o09050006 and 202423k09020003), and the Project of the Provincial Major Industrial Innovation Plan of Anhui Provincial Development and Reform Commission (Grant No.AHZDCYCXJH-OC2025-08).

References

[Achlioptas *et al.*, 2020] Panos Achlioptas, Ahmed Abdelreheem, Fei Xia, Mohamed Elhoseiny, and Leonidas Guibas. Referit3d: Neural listeners for fine-grained 3d object identification in real-world scenes. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, pages 422–440, Cham, 2020. Springer International Publishing.

[Azuma *et al.*, 2022] Daichi Azuma, Taiki Miyanishi, Shuhei Kurita, and Motoaki Kawanabe. Scanqa: 3d question answering for spatial scene understanding, 2022.

[Bai *et al.*, 2025] Shuai Bai, Keqin Chen, Xuejing Liu, et al. Qwen2.5-vl technical report, 2025.

[Chen *et al.*, 2020] Dave Zhenyu Chen, Angel X. Chang, and Matthias Nießner. Scanrefer: 3d object localization in rgb-d scans using natural language, 2020.

[Chen *et al.*, 2024] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks, 2024.

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

[Dai *et al.*, 2017] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes, 2017.

[Feng *et al.*, 2025] 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. A survey of large language model-powered spatial intelligence across scales: Advances in embodied agents, smart cities, and earth science, 2025.

[Guan *et al.*, 2024] Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. Hallusionbench: An advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models, 2024.

[Guo *et al.*, 2025] Dong Guo, Faming Wu, Feida Zhu, et al. Seed1.5-vl technical report, 2025.

[Hong *et al.*, 2023] Yining Hong, Haoyu Zhen, Peihao Chen, Shuhong Zheng, Yilun Du, Zhenfang Chen, and Chuang Gan. 3d-llm: Injecting the 3d world into large language models, 2023.

[Kuhn, 1955] Harold W. Kuhn. The hungarian method for the assignment problem. *Naval Research Logistics (NRL)*, 52, 1955.

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

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

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

[Liu *et al.*, 2025] Yuhong Liu, Beichen Zhang, Yuhang Zang, Yuhang Cao, Long Xing, Xiaoyi Dong, Haodong Duan, Dahua Lin, and Jiaqi Wang. Spatial-ssrl: Enhancing spatial understanding via self-supervised reinforcement learning, 2025.

[Ma *et al.*, 2023] Xiaojian Ma, Silong Yong, Zilong Zheng, Qing Li, Yitao Liang, Song-Chun Zhu, and Siyuan Huang. Sqa3d: Situated question answering in 3d scenes, 2023.

[Miao *et al.*, 2025] Xingyu Miao, Haoran Duan, Quanhao Qian, Jiuniu Wang, Yang Long, Ling Shao, Deli Zhao, Ran Xu, and Gongjie Zhang. Towards scalable spatial intelligence via 2d-to-3d data lifting, 2025.

[Nejatishahidin *et al.*, 2024] Negar Nejatishahidin, Madhukar Reddy Vongala, and Jana Kosecka. Structured spatial reasoning with open vocabulary object detectors, 2024.

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

[QwenLM Team, 2025] QwenLM Team. Qwen3-vl: Multimodal large language model series. https://github.com/QwenLM/Qwen3-VL, 2025. Accessed: 2025-11-14.

[Ravi *et al.*, 2020] Nikhila Ravi, Jeremy Reizenstein, David Novotny, Taylor Gordon, Wan-Yen Lo, Justin Johnson, and Georgia Gkioxari. Accelerating 3d deep learning with pytorch3d, 2020.

[Ravi *et al.*, 2024] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, et al. Sam 2: Segment anything in images and videos, 2024.

[Raychaudhuri and Chang, 2025] Sonia Raychaudhuri and Angel X. Chang. Semantic mapping in indoor embodied ai – a survey on advances, challenges, and future directions, 2025.

[Singh *et al.*, 2025] Aaditya Singh, Adam Fry, Adam Perelman, et al. Openai gpt-5 system card, 2025.

[Team *et al.*, 2025] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, et al. Gemini: A family of highly capable multimodal models, 2025.

[Wald *et al.*, 2020] Johanna Wald, Helisa Dhamo, Nassir Navab, and Federico Tombari. Learning 3D Semantic Scene Graphs from 3D Indoor Reconstructions. In *Proceedings IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[Wang *et al.*, 2023] Tai Wang, Xiaohan Mao, Chenming Zhu, Runsen Xu, Ruiyuan Lyu, Peisen Li, Xiao Chen, Wenwei Zhang, Kai Chen, Tianfan Xue, Xihui Liu, Cewu Lu, Dahua Lin, and Jiangmiao Pang. Embodiedscan: A holistic multi-modal 3d perception suite towards embodied ai, 2023.

[Wang *et al.*, 2024] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20697–20709, 2024.

[Wang *et al.*, 2025a] Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

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

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

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

[xAI, 2025] xAI. Grok 4. https://x.ai/news/grok-4, July 2025. Model announcement.

[Xu *et al.*, 2025] Qingyao Xu, Siheng Chen, Guang Chen, Yanfeng Wang, and Ya Zhang. Chatbev: A visual language model that understands bev maps, 2025.

[Yang *et al.*, 2023] Zhengyuan Yang, Linjie Li, Kevin Lin, Jianfeng Wang, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. The dawn of lmms: Preliminary explorations with gpt-4v(ision), 2023.

[Yang *et al.*, 2025a] Ganlin Yang, Tianyi Zhang, Haoran Hao, Weiyun Wang, Yibin Liu, Dehui Wang, Guanzhou Chen, Zijian Cai, Junting Chen, Weijie Su, Wengang Zhou, Yu Qiao, Jifeng Dai, Jiangmiao Pang, Gen Luo, Wenhai Wang, Yao Mu, and Zhi Hou. Vlaser: Vision-language-action model with synergistic embodied reasoning, 2025.

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

[Yang *et al.*, 2025c] Shusheng Yang, Jihan Yang, Pinzhi Huang, Ellis Brown, Zihao Yang, Yue Yu, Shengbang Tong, Zihan Zheng, Yifan Xu, Muhan Wang, Daohan Lu, Rob Fergus, Yann LeCun, Li Fei-Fei, and Saining Xie. Cambrian-s: Towards spatial supersensing in video, 2025.

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

[Yin *et al.*, 2025] Baiqiao Yin, Qineng Wang, Pingyue Zhang, Jianshu Zhang, Kangrui Wang, Zihan Wang, Jieyu Zhang, Keshigeyan Chandrasegaran, Han Liu, Ranjay Krishna, Saining Xie, Manling Li, Jiajun Wu, and Li Fei-Fei. Spatial mental modeling from limited views, 2025.

[Zhou *et al.*, 2019] Yang Zhou, Zachary While, and Evangelos Kalogerakis. Scenegraphnet: Neural message passing for 3d indoor scene augmentation, 2019.

[Zhu *et al.*, 2025] Jinguo Zhu, Weiyun Wang, Zhe Chen, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models, 2025.