CabinSI: Omni-Cabin Spatial Reasoning through Explicit Visual Cognitive Maps

Mengxue Qu[1,2⋆], Hengrui Hu[3], Mingming Ma[4], Ming Lei[4], Jie Gao[4], Henghui Ding[3], Yao Zhao[1,2], Kenn Wu[4], and Yunchao Wei[1,2]

[1] Institute of Information Science, Beijing Jiaotong University, China
[2] Visual Intelligence + X International Joint Laboratory of the Ministry of Education, China
[3] Institute of Big Data, College of Computer Science and Artificial Intelligence, Fudan University, China [4] NIO, China
{qumengxue,yunchao.wei,yzhao}@bjtu.edu.cn

Abstract. Intelligent cabin is gradually evolving from passive voice-based interfaces into multimodal collaborative systems supported by multi-camera and multi-sensor setups, where spatial reasoning becomes a core capability for enabling proactive perceptual interaction. Unlike traditional single-scene spatial tasks, cabin environments involve both interior and exterior vehicle regions, multiple occupants and traffic participants, and inherently require cross-view observations. To fill this research gap, we introduce *CabinSI*, a benchmark for evaluating spatial intelligence in cross-cabin environments, consisting of two components: RelCabin for relational reasoning tasks and RefCabin for spatial referring localization tasks. The benchmark is built upon real-world captured multi-view cabin data and systematically covers in-cabin, out-of-cabin, and cross-cabin scenarios under single-view, multi-view, and cross-cabin-view settings. Furthermore, we propose a cognitive-map-based framework that projects multi-view observations onto a normalized top-down plane and constructs an explicit spatial graph as model input, allowing multimodal large language models to focus on structured reasoning. Experimental results demonstrate that such explicit spatial representation significantly improves the stability of cross-view reasoning in MLLMs. Data and code is available at CabinSI.

1 Introduction

Traditional intelligent cabin primarily rely on passive voice-based interactions [36, 42]. As intelligent cabins evolve toward multimodal collaborative understanding, the in-vehicle environment is no longer isolated voice input, but a distributed observation system composed of multi-camera and multi-sensor setups. This system encompasses multi-source information including in-cabin occupants, interaction

⋆ Work done during an internship at NIO.

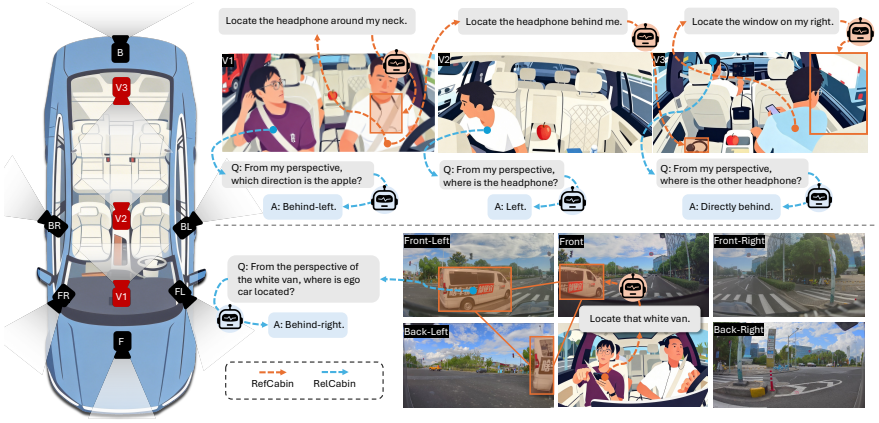

Fig. 1: Illustration of the CabinSI benchmark. The left panel shows the multi-camera configuration, including three in-cabin views (V1–V3) and six exterior views (F: front, FL: front-left, FR: front-right, BL: back-left, BR: back-right, B: back). The right panel presents examples of RelCabin and RefCabin tasks across in-cabin, out-of-cabin, and cross-cabin scenarios.

behaviors, and external dynamic environments, making proactive interaction initiation based on occupant states possible. In this scenario, many critical problems inherently depend on cross-view information integration and relational reasoning under a unified spatial reference frame. For example, determining relative orientations among passengers, understanding spatial associations between in-cabin actions and external events, or making joint decisions during interior-exterior interactions. Such problems extend beyond traditional single-image detection or question-answering paradigms, requiring models to construct consistent spatial cognitive structures under multi-view observations and perform cross-region, cross-object relational reasoning upon this foundation.

Existing spatial reasoning research primarily focuses on indoor or general scenarios, such as visual question answering [3, 27, 44, 62], referring expression comprehension [12, 24, 61], and embodied navigation tasks [2, 7, 34, 59], with scenes typically centered on households [2, 26, 67], autonomous driving environments [11, 29, 40], or simulated settings [25, 41]. Currently, there is no spatial reasoning research addressing the full-cabin environment. Cabin space is divided into interior and exterior regions, requiring the establishment of unified references across different physical spaces. Furthermore, scenes typically involve multiple occupants and external traffic participants, necessitating judgments of relative positions and interaction relationships among multiple objects. These characteristics render traditional spatial reasoning methods designed for single-scene or single-view settings [9, 52, 63] difficult to adapt directly, urgently calling for systematic modeling and evaluation frameworks specifically oriented toward cabin structural properties.

To systematically evaluate spatial reasoning capabilities in intelligent cabin scenarios, we construct ***CabinSI***, a **S**patial **I**ntelligence evaluation system for **Cabin** environments. CabinSI comprises two components: RefCabin evaluates referring recognition capabilities across in-cabin, out-of-cabin, and cross-cabin scenarios; RelCabin evaluates relative spatial relationship reasoning under a unified spatial reference frame. Overall tasks cover three spatial scopes (in-cabin, out-of-cabin, and cross-cabin) and include three input modalities (single-view, multi-view, and cross-cabin-view), thereby systematically characterizing spatial consistency modeling in multi-camera environments. Unlike traditional evaluation settings centered on detection accuracy or single-image question answering [48, 57, 68], CabinSI emphasizes: (1) evaluation paradigms covering in-cabin, out-of-cabin, and cross-cabin spaces; (2) spatial evaluation under single-view, multi-view, and cross-cabin-view configurations; (3) dual-capability assessment of spatial relational reasoning and referring localization. Through explicit task division and evaluation protocols, CabinSI transforms spatial understanding capabilities in cabin scenarios from implicit manifestations into decomposable, quantifiable, and comparable research problems.

Furthermore, we propose an explicit-cognitive-map-based framework to address the bottleneck of multi-view cross-cabin spatial reasoning. We first leverage detection and depth information to project multi-view observations from both interior and exterior regions onto a normalized top-down plane. Under the front-back-left-right reference system, we construct an explicit spatial graph composed of object nodes. Serving as an intermediate representation between perception and multimodal large models [22,45,60], this map supports reference-centered coordinate transformation and geometric partitioning for RelCabin, enabling stable judgments of relative relations such as "front/back/left/right". Simultaneously, it transforms referring expression comprehension in RefCabin into explicit node selection over the map, achieving consistent cross-view, cross-region localization through joint matching of semantic attributes and spatial constraints. Our experiments demonstrate that the proposed Cognitive Map framework substantially improves spatial reasoning stability across different model scales and viewpoint configurations.

In summary, our contributions are threefold:

- We formally define the cross-cabin spatial reasoning problem and systematically analyze spatial modeling challenges under multi-view and coupled interior-exterior scenarios.
- We construct the CabinSI benchmark, achieving systematic evaluation of cabin spatial localization and relational reasoning capabilities through RefCabin and RelCabin tasks.
- We propose a structured spatial representation method based on cognitive maps and validate the critical role of proper spatial representation design in stabilizing multimodal model spatial reasoning.

Table 1: Comparison of datasets/benchmarks. **MI**: multi-image; **SR**: spatial relation; **MI∩SR**: multi-image with explicit spatial relations; **I/O**: in/out-of-car setting. ✔ denotes partial support (e.g., only a subset of samples/tasks).

Dataset	Venue	Samples	QA	Grd.	Drive	MI	MI∩SR	SR	I/O
Autonomous Driving									
Talk2Car [11]	EMNLP'19	9.2K	✗	✔	✔	✗	✗	✗	✗
DriveLM [44]	ECCV'24	4.1K	✔	✔	✔	✗	✗	✗	✗
Nuscenes-QA [40]	AAAI'24	34.1K	✔	✗	✔	✔	✔	✔	✗
Nuprompt [56]	AAAI'25	34.1K	✗	✔	✔	✔	✔	✔	✗
NuInteract [64]	arXiv'25	34.1K	✔	✔	✔	✔	✔	✔	✗
General Multi-Image Understanding									
MMMU [62]	CVPR'24	11.5K	✔	✗	✗	✔	✗	✗	✗
LLaVA-Interleave [28]	ICLR'25	17.0K	✔	✗	✔	✔	✗	✔	✗
MC-Bench [58]	CVPR'25	2.0K	✔	✔	✗	✔	✔	✔	✗
Migician [30]	ACL'25	530.0K	✗	✔	✗	✔	✔	✔	✗
Spatial Reasoning									
Space3D-Bench [47]	ECCV'24	1.0K	✔	✗	✗	✗	✗	✔	✗
VSI-Bench [60]	CVPR'25	3.7K	✔	✗	✗	✔	✗	✔	✗
Ego3D-Bench [17]	ICLR'26	8.6K	✔	✔	✔	✔	✔	✔	✗
Spatialab [54]	arXiv'26	1.4K	✔	✗	✔	✗	✗	✔	✗
CabinSI (Ours)	–	3.7K	✔	✔	✔	✔	✔	✔	✔

2 Related Work

2.1 Driving Scene Datasets and Benchmarks

In the context of autonomous driving and intelligent transportation, numerous datasets have been proposed to evaluate visual question answering and referring localization capabilities. For instance, nuScenes-QA [40] and DriveLM [44] focus on question answering and reasoning in outdoor traffic environments. Datasets such as Talk2Car [11], DRAMA [35], and NuPrompt [56] emphasize natural language grounding of traffic participants. Recent efforts, including NuScenes-MQA [19] and NuInteract [64], extend evaluation to multi-image or interactive question answering settings. Furthermore, most recently advancements like VLAAD [37] and OmniDrive [53] bridge vision-language modeling with driving action sequences, further pushing the boundaries of scene understanding. These benchmarks have significantly advanced visual understanding and decision modeling in outdoor driving scenarios. However, they primarily center on external perception of out-cabin regions, focusing on object detection, lane topology, and macro-traffic dynamics. They overlook the in-cabin occupants and fail to evaluate the joint spatial reasoning across interior and exterior regions. Even when multiple images are supported, the spatial scope remains confined to exterior, lacking systematic evaluation of cross-cabin spatial consistency modeling.

2.2 General Multi-Image and Multimodal Benchmarks

With the rapid evolution of Multimodal Large Language Models (MLLMs), various benchmarks have emerged to evaluate their capabilities across interleaved or multi-image inputs. Foundational benchmarks such as MMBench [33] and MMMU [62] focus on domain-specific knowledge and reasoning, while more recent efforts like LLaVA-Interleave [28], MANTIS [21], and MileBench [13] emphasize long-context processing and information fusion across multiple images. MUIRBENCH [50], MC-Bench [58] and Migician [30] propose more complex tasks, including fundamental visual perception and complex reasoning over multiple visual contexts. While advancing multi-image reasoning, these benchmarks primarily operate in 2D semantic or temporal domains, lacking rigorous 3D geometric constraints or unified spatial reference frames across views. Images are often semantically related rather than geometrically aligned, thus the integration does not inherently translate to cross-view spatial consistency. Existing general benchmarks therefore provide limited insight into whether models can construct coherent spatial representations in real-world distributed physical spaces.

2.3 Spatial Reasoning Benchmarks

Spatial reasoning, the ability to infer geometric structure and relational layout from visual inputs, is central to visual intelligence. Beyond early synthetic or templated benchmarks like CLEVR [23] and GQA [18], recent evaluations shift toward naturalistic settings and explicitly target spatial failures. ScanQA [4], EmbSpatial-Bench [14], and Space3D-Bench [47] focus on 3D-grounded QA and multi-step spatial inference. Some benchmarks, such as Visual Spatial [31], BLINK-Spatial [16], and VSI-Bench [60], center on diagnostic tests of core spatial primitives such as relative direction, occlusion, and depth ordering. Moreover, SpatialVLM [8], Spatial-MM [43], and OmniSpatial [20] emerge as large-coverage, open-form benchmarks. Most recently more application- and time-aware settings are introduced by VLM4D [65], SpatialRGPT-Bench [10], Mind the Gap [46], SPATIALAB [54], and Struct2D [66], etc. Despite these advances, current benchmarks often simplify the problem by operating within a single region or treating regions as loosely related, and even when multiple views are available they rarely enforce strict cross-region reference-frame consistency. This leaves structurally coupled interior–exterior scenarios underexplored, where success requires coherent geometry across regions and stable relational reasoning under dynamic interactions. CabinSI is designed to address this gap with a quantifiable evaluation centered on cross-cabin consistency and multi-view integration.

3 CabinSI-Bench

3.1 Problem Definition

Most existing autonomous driving benchmarks focus on external traffic environments and implicitly assume a single external spatial region. In real-world

intelligent cabin scenarios, a model must simultaneously reason over **interior (S_{in}) and exterior (S_{out}) regions**, integrating observations from **distributed multi-camera systems** ($\mathcal{I} = \{I^1, \ldots, I^V\}$) and establishing a **unified spatial reference frame** (R). This requires models to construct a shared spatial representation $f(\mathcal{I}, R)$ such that relational queries across regions or objects can be reliably answered. Current benchmarks do not explicitly evaluate this cross-region, multi-view reasoning capability.

To address this, we formalize **Cross-Cabin Spatial Reasoning** as follows: given a set of multi-view images ($\mathcal{I}$) and a language query (q), the model must output either a set of localized coordinates ($\mathcal{L} = \{l_1, \ldots, l_m\}$) for referred entities or a textual answer (a) to a relational query, such that:

$$f(\mathcal{I}, q; R) \to \mathcal{L} \quad \text{or} \quad f(\mathcal{I}, q; R) \to a \tag{1}$$

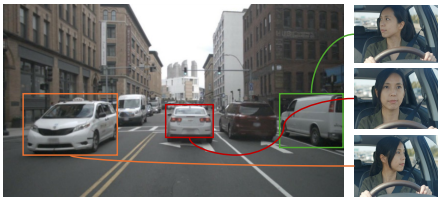

Fig. 2: Example of cross-cabin reasoning. The occupant's gesture and gaze help disambiguate which white car is the intended target when multiple vehicles are visible. Faces are AI-generated schematics.

This formulation emphasizes **cross-region, cross-view, and multi-object spatial reasoning** under a shared reference frame, abstracting away from sensor-specific details while capturing the core reasoning challenge. As illustrated in Figure 2, when dealing with unclear instructions such as *"that white car"* while multiple cars are visible in front, the task cannot be reliably solved using only text and exterior cameras. Additional contextual cues such as the gesture, head orientation, or gaze direction are required to correctly interpret the instruction.

CabinSI-Bench operationalizes this problem through two complementary task families. **RefCabin** requires localizing entities given a referring expression, evaluating multi-view grounding and spatial consistency. **RelCabin** focuses on explicit relational reasoning, where the answers to queries require understanding relative positions and interactions across interior and exterior regions.

3.2 Data Collection

CabinSI-Bench is built from synchronized multi-camera captures in real-world driving scenarios and carefully curated language commands (Fig. 3). The dataset includes two subsets, RelCabin and RefCabin, covering both in-cabin and out-of-cabin viewpoints. It combines portions of existing multi-modal autonomous driving datasets with newly collected in-cabin and cross-cabin data.

In-cabin data. Based on our self-collected single-view and multi-view in-cabin images, we annotated a batch of referring expression grounding data, which was incorporated into RefCabin. Subsequently, we utilized GPT [10] to rewrite these into spatial reasoning QA data, which was incorporated into RelCabin.

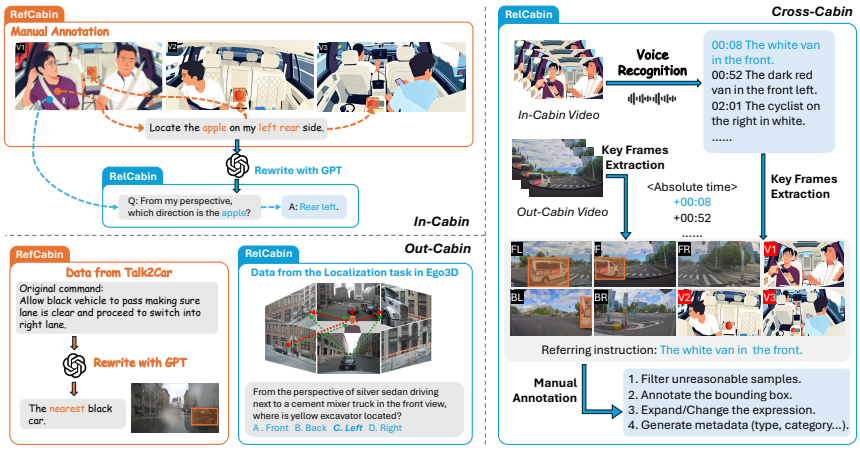

Fig. 3: Overview of the data construction pipeline in CabinSI-Bench. The figure illustrates three main scenarios: in-cabin (top-left), out-of-cabin (bottom-left), and cross-cabin (right), showing how multimodal streams and existing driving benchmarks are transformed into RefCabin and RelCabin samples.

Out-of-cabin data. This part is constructed from the Ego3D [17] spatial reasoning dataset and the referring expression grounding dataset Talk2Car [11], both of which provide rich natural language descriptions of traffic participants. For Talk2Car, we rewrote complex original commands (e.g., "Allow the black vehicle to pass, ensure the lane is clear, and then proceed to switch into the right lane.") into concise spatial questions using GPT, which were incorporated into RefCabin. For Ego3D, its Localization subset perfectly matches the problem types addressed in this work, so we directly incorporated it into RelCabin.

Cross-cabin data. We additionally collected synchronized interior–exterior multi-camera sequences using an instrumented vehicle, designed to elicit cross-cabin spatial relations under a unified reference frame. Our platform employs 11 cameras in total, including 8 exterior cameras and 3 interior cameras. As shown in Fig. 1, the exterior system covers front, rear, front-left, front-right, rear-left, and rear-right views to increase coverage of near-field regions around the vehicle. The interior system consists of three viewpoints: a front-facing camera oriented rearward, a roof-mounted camera oriented rearward, and a rear-roof camera oriented forward, jointly providing complementary observations of occupants and cabin layout. During data collection, occupants naturally issue spoken commands (e.g., "Locate the apple on my left rear side."), which are transcribed by a speech recognition system and aligned with video streams using absolute timestamps. For each utterance, we perform key-frame extraction around the corresponding time and crop synchronized multi-view images from interior and exterior camera systems to form the visual input set, as shown in Fig. 3.

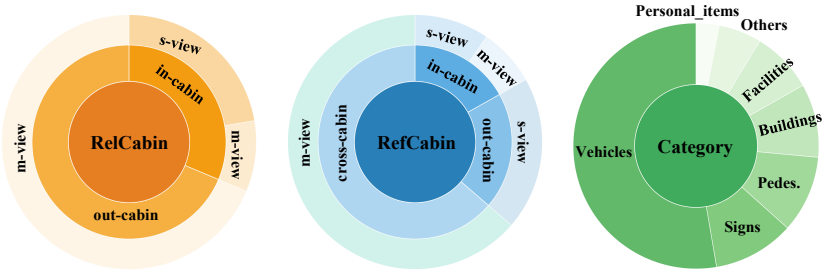

Fig. 4: Statistical overview of CabinSI-Bench: (a) RelCabin and (b) RefCabin sample distributions, and (c) object category distribution for cross-cabin samples.

Manual annotation and quality control. For RefCabin, annotators first filter unreasonable or ambiguous utterances, then draw bounding boxes for referred objects on selected views. They further expand or refine the expressions when necessary (e.g., clarifying view or side) and generate structured metadata (category, target type, view IDs, cabin region, etc.) to support downstream analysis. For RelCabin, annotators validate the GPT-rewritten questions, correct or discard ambiguous cases, and confirm the final front/back/left/right labels. Across both tasks, all samples are curated under the following principles: each query involves **explicit spatial relationships**, cross-cabin samples require information from both interior and exterior regions, and multi-view samples require view integration rather than single-image reasoning. All samples undergo multi-round human verification to ensure semantic clarity, answer uniqueness, and spatial logical consistency.

3.3 Dataset Statistics

CabinSI-Bench contains **3,758 samples** in total, with **1,121 RelCabin** relational reasoning samples and **2,637 RefCabin** referring grounding samples. As shown in Fig. 4, the dataset spans **in-cabin**, **out-of-cabin**, and **cross-cabin** spatial scopes, with both **single-view** and **multi-view** configurations. RelCabin covers in-cabin scenarios (250 single-view, 101 multi-view) and out-of-cabin multi-view scenarios (770 samples). RefCabin includes 250 in-cabin single-view, 194 in-cabin multi-view samples, 514 out-cabin single-view scenarios, and 1679 cross-cabin multi-view scenarios, enabling comprehensive evaluation of cross-region spatial grounding.

The category distribution of cross-cabin samples is prominently led by Vehicles (52.7%). Traffic Signs (10.7%) and Pedestrians (10.2%) constitute the primary dynamic categories, while Buildings (9.7%) and Facilities (8.1%) define the structural environment. The Personal Items (3.1%) that appear inside the cabin are also included. The diverse composition ensures a comprehensive representation of entities within the operational domain.

Overall, CabinSI-Bench represents the first systematic benchmark for evaluating **cross-cabin, multi-view spatial intelligence**. By explicitly modeling

coupled interior–exterior regions and enforcing multi-view spatial consistency, it provides a unified platform for assessing models' abilities in **explicit spatial reasoning, cross-region relational inference, and multi-object grounding** in realistic cabin environments.

4 Method

4.1 Cognitive Map Creation

Following Ego3D-VLM [17], we treat cognitive map construction as a post-training procedure that converts ego-centric multi-view observations into a compact, ego-centered world model. Given synchronized multi-view images $\{I^v\}_{v=1}^{V}$ and a natural-language question q, a referring expression comprehension (REC) model is first applied independently to each view to detect all objects mentioned in q. For each camera v, the REC model predicts a 2D bounding box $b_i^{(v)}$ corresponding to each related object $o_i^{(v)}$ in q, yielding a set of matched pairs $\mathcal{O}^{(v)} = \{(o_i^{(v)}, b_i^{(v)})\}_{i=1}^{N^{(v)}}$. From these predicted boxes, we compute the 2D pixel-space centers $u_i^{(v)} \in \mathbb{R}^2$ for all detected objects.

To lift these detections into 3D, Ego3D-VLM employs a metric depth estimator that predicts a dense depth map $D^{(v)}$ for each view. For every object center $u_i^{(v)} = (x_i, y_i)$, it extracts the corresponding depth value $d_i^{(v)} = D^{(v)}(x_i, y_i)$ and projects the point into the camera coordinate system with intrinsic matrix $K^{(v)}$:

$$p_{\text{cam},i}^{(v)} = d_i^{(v)} (K^{(v)})^{-1} [x_i, y_i, 1]^\top. \tag{2}$$

With known or estimated extrinsics $(R^{(v)}, T^{(v)})$, all camera-centric 3D points are then transformed into a unified global coordinate system, yielding ego-centric positions $p_{\text{global},i}^{(v)}$. A relational scaling step further normalizes these coordinates using canonical object sizes (e.g., pedestrians, sedans) to obtain physically plausible distances even under imperfect metric depth.

Finally, Ego3D-VLM defines a cognitive map generator F_{cog} that converts the set of global 3D coordinates and their associated expressions into a textual cognitive map, $C = F_{\text{cog}}\big(\{(o_i^{(v)}, p_{\text{global},i}^{(v)})\}_{i,v}\big)$. This textual cognitive map summarizes, in an ego-centric reference frame, which objects are present, where they are located in 3D space, and from which viewpoints they are observed. When fed jointly with the original multi-view images and the question to a VLM, this structured spatial representation substantially improves 3D spatial reasoning performance on ego-centric multi-view benchmarks. In CabinSI-Bench, we further adapt this general formulation into an explicit, view-invariant 2D layout tailored to our RelCabin and RefCabin tasks.

4.2 Explicit Cognitive Map Framework for RelCabin

Building on the above cognitive map creation pipeline, we adopt an explicit cognitive map paradigm to represent ego-centric multi-view scenes with a view-invariant spatial layout for RelCabin, as shown in Fig. 5. To effectively construct

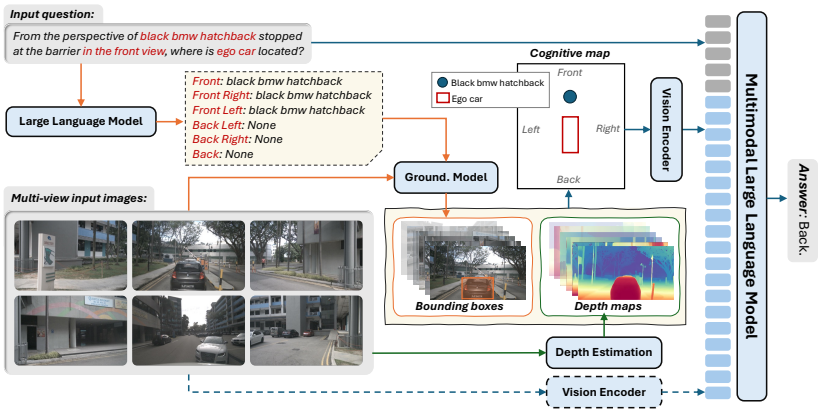

Fig. 5: Overview of the explicit cognitive map framework for RelCabin. Given a natural language query and synchronized multi-view images, an LLM first pre-parses the query to select relevant views and objects. Detected objects are projected into a unified top-down coordinate system to construct the CogMap, enabling ego-centric relational reasoning. The dotted arrow indicates that inputting images is optional.

and utilize cognitive maps in CabinSI-Bench, we make two key design choices: (i) before map construction, we explicitly pre-parse the query to select exclusively the views and objects that are semantically required, and (ii) downstream models are required to answer solely from the rendered CogMap image, without access to textual encodings of the map. This design aims to reduce redundancy in the map, tighten the coupling between language and spatial abstraction, and push cognitive maps toward a form that can ultimately be produced by image generation tools rather than precise 3D reconstruction.

Formally, given a natural language query q and synchronized multi-view inputs $\mathcal{I} = \{I^v\}$, we first leverage an LLM to pre-interpret q, identify the involved object categories and filter the subset of views that are likely to contain relevant evidence. The Cognitive Map is then defined as an explicit spatial graph $\mathcal{M} = \{(o_i, x_i, y_i)\}_{i=1}^{N}$, where each node corresponds to a semantically grounded object instance and (x_i, y_i) denotes its position in a shared 2D top-down coordinate system. Unlike naive fusion strategies that simply concatenate multi-view features, our map construction enforces geometric consistency by projecting detected objects into a shared spatial frame, while *deliberately* targeting an approximate 2D layout rather than metrically precise 3D geometry. Accurate 3D reconstruction typically requires dense sensor rigs and careful calibration, whereas approximate top-down layouts can be sketched by humans or potentially generated from ego views by image models [15, 39, 55].

In our pipeline, the final MLLM receives the original question q together with a rendered CogMap image, but *no* textual serialization of $\mathcal{M}$. That is, spatial structure is externalized into a manipulable top-down sketch, and the MLLM reasons over this image instead of directly over raw cabin views. This brings

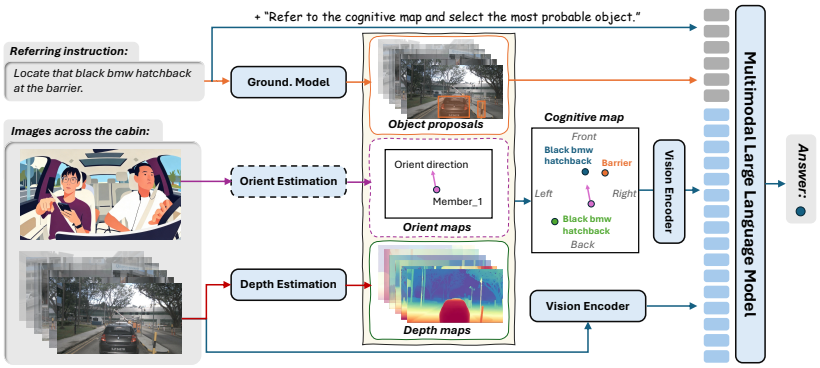

Fig. 6: Overview of the explicit cognitive map framework for RefCabin. Given a referring instruction and synchronized multi-view observations, we build a unified top-down cognitive map with labeled object nodes and resolve the reference by selecting the most compatible node on this map. The dotted box "Orient Estimation" indicates an optional component, with direction prediction serving as an optional input.

two advantages: first, it isolates spatial reasoning from low-level appearance, enabling principled coordinate transformation, reference-centered reasoning, and relational constraint evaluation; second, it opens a practical path toward future systems where CogMaps are produced directly by generative models (or human sketches) in the top-down space, bypassing the need for expensive 3D reconstruction while still supporting RelCabin and RefCabin style spatial reasoning.

4.3 Explicit Cognitive Map Framework for RefCabin

RefCabin addresses cross-cabin referring expression grounding, where a natural language instruction specifies a target object potentially described through attributes and spatial constraints. Given multi-view observations $\mathcal{I}$ and a referring expression q_{ref}, the goal is to identify the target object instance o^*. Unlike conventional single-view referring tasks, RefCabin must resolve ambiguities arising from distributed viewpoints and partial visibility across cameras.

We construct the Cognitive Map as described above, projecting object proposals from all views into a unified top-down coordinate system, as shown in Fig. 6. The referring instruction is decomposed into semantic attributes (e.g., category, color, identity) and relational constraints (e.g., "at the barrier", "next to the ego car"). These constraints are evaluated against the geometric and topological properties encoded in the map.

Referring resolution is formulated as an explicit node selection problem:

$$o^* = \arg\max_{o_i \in \mathcal{M}} f_{sel}(o_i, q_{ref}), \tag{3}$$

where f jointly measures semantic compatibility and spatial consistency. Spatial phrases are translated into geometric predicates, such as proximity thresholds or

relative ordering constraints in the coordinate space. Because the Cognitive Map encodes explicit object positions, relational grounding becomes a deterministic verification process rather than a soft attention alignment over visual tokens.

Collectively, the cognitive map framework for RelCabin and RefCabin embodies a unified principle: spatial intelligence in multi-view cabin scenarios arises from explicit geometric representations and reasoning, rather than from end-to-end implicit feature fusion. In this context, the Cognitive Map acts as a computational abstraction layer that bridges the gap between perception and reasoning, facilitating perspective transformation, relational verification, and cross-view disambiguation within a consistent spatial coordinate system.

5 Experiments

We conduct comprehensive experiments on RelCabin and RefCabin to evaluate the effectiveness of the proposed cognitive map framework for in-out-cross cabin spatial reasoning and grounding. The experiments compare multiple backbone MLLMs under single-view and multi-view settings, and include controlled ablations to analyze the role of viewpoint selection, question-guided object identification, and cognitive map construction. We further investigate how the accuracy of the intermediate spatial representation influences final reasoning performance through additional studies.

5.1 Experimental Setup

Backbone Models. We evaluate our method on multiple multimodal large language models: GPT-5.2 [1], Gemini-3-Flash [49], Qwen2.5-VL-7B [6], and Qwen3-VL-4B/8B [5].

Evaluation Metrics. RelCabin has two different question types: (1) **QA**: the model answers with directional phrases; we use an LLM to compare the model output with the ground-truth answer and treat the sample as correct when the two are semantically consistent; accuracy is the fraction of correct QA samples. (2) **Multiple choice**: the model selects one option; a sample is correct if the selected option matches the ground truth; we report accuracy over these samples. For RefCabin, we feed multi-view images and take the model's predicted bounding box; a prediction is correct if its IoU with *any* ground-truth box exceeds 0.5. We report localization accuracy as the proportion of correct predictions.

5.2 Main Results on RelCabin

Tab. 2 presents the quantitative comparison on RelCabin, where we focus on the *Overall* accuracy aggregated over all subsets. Across all three backbone models, our Cognitive Map Framework consistently yields substantial overall gains: accuracy improves from 25.31% to 41.17% (+15.86), from 27.17% to 44.46% (+17.29), and from 33.59% to 45.92% (+12.33) for Qwen2.5-VL-7B, Qwen3-VL-4B, and

Table 2: Main results on RelCabin. s_view and m_view denote single-view and multi-view settings. All results are reported in percentages.

Model	In-Cabin		Out-Cabin	Overall
	s_view	m_view	m_view	
Qwen2.5-VL-7B	14.00	16.83	30.09	25.31
+ **Ours**	**27.64**	**26.73**	**47.46**	**41.17**
Qwen3-VL-4B	13.60	18.81	32.68	27.17
+ **Ours**	**25.71**	**30.69**	**52.36**	**44.46**
Qwen3-VL-8B	10.40	19.80	42.93	33.59
+ **Ours**	**29.39**	**24.75**	**54.07**	**45.92**

Qwen3-VL-8B, respectively. These results demonstrate that explicit spatial abstraction markedly enhances the models' reasoning capability on RelCabin.

Interestingly, Qwen3-VL models occasionally underperform Qwen2.5-VL models on certain subsets, likely because they tend to over-deliberate when the Cognitive Map contains minor inaccuracies.

5.3 Ablation Study on RelCabin

Tab. 3 presents ablation results on out-cabin multi-view RelCabin, revealing two key factors for effective spatial reasoning.

First, **view selection and semantic filtering** are critical for performance. Comparing (3) and (4), filtering for the target view yields a marginal improvement (32.99% vs. 33.15%), suggesting that superfluous views can introduce confounding noise. More notably, replacing heuristic rule-based labels with GPT-extracted relevant objects (4, 7 vs. 8) results in a substantial performance surge. Specifically, the transition from GPT-based parsing (47.46%) compared to rule-based methods (37.48%) underscores that the accurate identification of question-relevant entities is essential for effective spatial reasoning.

Second, **a clean cognitive map representation** outperforms raw visual inputs. When only the cognitive map image is provided without raw images (5, 8), performance reaches 37.48% and 47.46% respectively, surpassing all variants with multi-view image inputs (from 30.09% to 35.06%). This confirms that the cognitive map provides sufficient spatial information while eliminating distracting low-level details, enabling focused reasoning on spatial relations rather than pixel-level appearance.

In summary, the performance gains stem from constructing a **semantically filtered, view-selected, and spatially abstracted** cognitive map, not from simply stacking detection modules or retaining all visual information.

5.4 Main Results on RefCabin

Tab. 4 presents RefCabin results. Compared with direct bounding box prediction, introducing proposal generation improves multi-view grounding performance significantly (from 14.95% to 22.16% in in-cabin multi-view).

Table 3: Ablation study on out-cabin multi-view RelCabin.

Method	Image Input	Grounding Input	CogMap Form	Acc
Baseline	Multi-view	None	None	30.09
(1)	All Multi-view	Sentence	Text	31.78
(2)	All Multi-view	Rule-extracted	Text+Image	31.56
(3)	All Multi-view	Rule-extracted	Image	32.99
(4)	Only Target-view	Rule-extracted	Image	33.15
(5)	None	Rule-extracted	Image	37.48
(6)	Only Target-view	GPT-extracted	Text+Image	33.20
(7)	Only Target-view	GPT-extracted	Image	35.06
(8)	None	GPT-extracted	Image	**47.46**

Table 4: RefCabin grounding results using Qwen2.5-VL-7B backbone. ✔ and ✗ indicate whether proposal generation, cognitive map construction, and selection-based inference are used.

Method	Prop.	CogMap	Sel.	In-Cabin		Out-Cabin		Cross-Cabin	Overall
				s	m	s	m	(m)	
(1)	✗	✗	✗	49.41	14.95	55.64	30.61	–	–
(2)	✔	✗	✔	48.26	22.16	58.37	35.97	29.32	39.48
(3)	✔	✔	✔	**52.90**	**24.23**	**59.14**	**36.02**	**30.33**	**40.40**

Further introducing cognitive map construction consistently yields the best performance across all settings, achieving 52.90%, 24.23%, 59.14%, and 36.02% on different subsets.

The improvement is especially pronounced in multi-view scenarios, confirming that cross-view grounding fundamentally requires spatial alignment rather than independent detection ranking.

5.5 Effect of Cognitive Map Quality

To investigate the upper bound of our framework, we conduct an experiment using human-annotated cognitive maps on a subset of 100 examples. Specifically, instead of relying on GroundingDINO [32] and depth estimation for 3D projection [38, 51], human annotators directly label object positions on the top-down cognitive map based on their perception of all views. Tab. 5 compares each model's performance with direct image input, predicted cognitive map, and human-annotated cognitive map.

The results reveal a clear capability gap between large and small models. With accurate cognitive maps, GPT-5.2 and Gemini achieve near-perfect performance (71–96%), while smaller models such as GPT-5-nano and Qwen2.5-VL show more limited gains (25–55%). More importantly, providing human-annotated maps dramatically improves performance across all models (e.g., GPT-5.2 from 12–38% to 74–96%), indicating that the main bottleneck lies in cognitive map prediction rather than the reasoning framework itself. These results suggest that learning to directly generate accurate cognitive maps from multi-view inputs could substantially improve spatial reasoning performance.

Table 5: Performance comparison using human-annotated cognitive maps. Each column is evaluated on 100 samples for the corresponding setting (300 samples in total). s_{view} means single-view, m_{view} means multi-view.

Model	In-Cabin		Out-Cabin	Overall
	s_{view}	m_{view}	m_{view}	
GPT-5.2	17	12	38	30.97
+ CogMap (Human)	**96**	**95**	**74**	80.80
Gemini-3-Flash	31	26	52	44.97
+ CogMap (Human)	**86**	**95**	**71**	76.51
GPT-5-nano	19	17	12	14.01
+ CogMap (Human)	**52**	**55**	**25**	33.72
Qwen2.5-VL-7B	20	18	27	24.63
+ CogMap (Predicted)	32	25	**50**	43.73
+ CogMap (Human)	**79**	**51**	38	48.31

5.6 Discussion

Although CabinSI focuses on the intelligent cabin scenario, the benchmark remains highly challenging due to diverse spatial layouts, object categories, and cross-view relations. Experiments show that current MLLMs struggle to form consistent spatial reasoning across views, particularly under occlusions, cluttered scenes, and cross-cabin settings. As such, CabinSI complements existing benchmarks by providing a realistic testbed for evaluating spatial intelligence in human–cabin interaction.

6 Conclusion

We present CabinSI, the first systematic benchmark for evaluating spatial intelligence in intelligent cabins, covering RefCabin and RelCabin tasks across in-cabin, out-of-cabin, and cross-cabin scenarios under single-view and multi-view settings. We further introduce an explicit cognitive map framework that enables MLLMs to perform spatial reasoning on structured spatial representations. Experiments show that cognitive maps substantially improve spatial reasoning, with ablations highlighting the importance of semantic filtering and spatial abstraction, while remaining limitations mainly stem from perception accuracy rather than the reasoning framework.

Acknowledgements This work was supported by the National Natural Science Foundation of China (No.92470203, No. 62472104).

References

1. Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al.: Gpt-4 technical report. arXiv preprint arXiv:2303.08774 (2023)
2. Anderson, P., Wu, Q., Teney, D., Bruce, J., Johnson, M., Sünderhauf, N., Reid, I., Gould, S., Van Den Hengel, A.: Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In: CVPR. pp. 3674–3683 (2018)
3. Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C.L., Parikh, D.: Vqa: Visual question answering. In: ICCV. pp. 2425–2433 (2015)
4. Azuma, D., Miyanishi, T., Kurita, S., Kawanabe, M.: Scanqa: 3d question answering for spatial scene understanding. In: CVPR. pp. 19129–19139 (2022)
5. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., et al.: Qwen3-vl technical report. arXiv preprint arXiv:2511.21631 (2025)
6. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., Zhong, H., Zhu, Y., Yang, M., Li, Z., Wan, J., Wang, P., Ding, W., Fu, Z., Xu, Y., Ye, J., Zhang, X., Xie, T., Cheng, Z., Zhang, H., Yang, Z., Xu, H., Lin, J.: Qwen2.5-vl technical report (2025), https://arxiv.org/abs/2502.13923
7. Cai, W., Huang, S., Cheng, G., Long, Y., Gao, P., Sun, C., Dong, H.: Bridging zero-shot object navigation and foundation models through pixel-guided navigation skill. In: ICRA. pp. 5228–5234. IEEE (2024)
8. Chen, B., Xu, Z., Kirmani, S., Ichter, B., Sadigh, D., Guibas, L., Xia, F.: Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. In: CVPR. pp. 14455–14465 (2024)
9. Chen, Y., Huang, S., Yuan, T., Qi, S., Zhu, Y., Zhu, S.C.: Holistic++ scene understanding: Single-view 3d holistic scene parsing and human pose estimation with human-object interaction and physical commonsense. In: ICCV. pp. 8648–8657 (2019)
10. Cheng, A.C., Yin, H., Fu, Y., Guo, Q., Yang, R., Kautz, J., Wang, X., Liu, S.: Spatialrgpt: Grounded spatial reasoning in vision-language models. NIPS **37**, 135062–135093 (2024)
11. Deruyttere, T., Vandenhende, S., Grujicic, D., Van Gool, L., Moens, M.F.: Talk2car: Taking control of your self-driving car. In: IJCNLP. pp. 2088–2098 (2019)
12. Ding, H., Liu, C., He, S., Jiang, X., Loy, C.C.: Mevis: A large-scale benchmark for video segmentation with motion expressions. In: ICCV. pp. 2694–2703 (2023)
13. Dingjie, S., Chen, S., Chen, G.H., Yu, F., Wan, X., Wang, B.: Milebench: Benchmarking mllms in long context. In: COLM (2024)
14. Du, M., Wu, B., Li, Z., Huang, X.J., Wei, Z.: Embspatial-bench: Benchmarking spatial understanding for embodied tasks with large vision-language models. In: ACL. pp. 346–355 (2024)
15. Esser, P., Kulal, S., Blattmann, A., Entezari, R., Müller, J., Saini, H., Levi, Y., Lorenz, D., Sauer, A., Boesel, F., et al.: Scaling rectified flow transformers for high-resolution image synthesis. In: ICLR (2024)
16. Fu, X., Hu, Y., Li, B., Feng, Y., Wang, H., Lin, X., Roth, D., Smith, N.A., Ma, W.C., Krishna, R.: Blink: Multimodal large language models can see but not perceive. In: ECCV. pp. 148–166. Springer (2024)
17. Gholami, M., Rezaei, A., Weimin, Z., Zhang, Y., Akbari, M.: Spatial reasoning with vision-language models in ego-centric multi-view scenes. ICLR (2026)

18. Hudson, D.A., Manning, C.D.: Gqa: A new dataset for real-world visual reasoning and compositional question answering. In: CVPR. pp. 6700–6709 (2019)
19. Inoue, Y., Yada, Y., Tanahashi, K., Yamaguchi, Y.: Nuscenes-mqa: Integrated evaluation of captions and qa for autonomous driving datasets using markup annotations. In: WACV. pp. 930–938 (2024)
20. Jia, M., Qi, Z., Zhang, S., Zhang, W., Yu, X., He, J., Wang, H., Yi, L.: Omnispatial: Towards comprehensive spatial reasoning benchmark for vision language models. arXiv preprint arXiv:2506.03135 (2025)
21. Jiang, D., He, X., Zeng, H., Wei, C., Ku, M., Liu, Q., Chen, W.: Mantis: Interleaved multi-image instruction tuning. arXiv preprint arXiv:2405.01483 (2024)
22. Jin, Y., Li, J., Gu, T., Liu, Y., Zhao, B., Lai, J., Gan, Z., Wang, Y., Wang, C., Tan, X., et al.: Efficient multimodal large language models: A survey. Visual Intelligence **3**(1), 27 (2025)
23. Johnson, J., Hariharan, B., Van Der Maaten, L., Fei-Fei, L., Lawrence Zitnick, C., Girshick, R.: Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In: CVPR. pp. 2901–2910 (2017)
24. Kazemzadeh, S., Ordonez, V., Matten, M., Berg, T.: Referitgame: Referring to objects in photographs of natural scenes. In: EMNLP. pp. 787–798 (2014)
25. Kolve, E., Mottaghi, R., Han, W., VanderBilt, E., Weihs, L., Herrasti, A., Deitke, M., Ehsani, K., Gordon, D., Zhu, Y., et al.: Ai2-thor: An interactive 3d environment for visual ai. arXiv preprint arXiv:1712.05474 (2017)
26. Ku, A., Anderson, P., Patel, R., Ie, E., Baldridge, J.: Room-across-room: Multilingual vision-and-language navigation with dense spatiotemporal grounding. In: EMNLP. pp. 4392–4412 (2020)
27. Kuang, J., Shen, Y., Xie, J., Luo, H., Xu, Z., Li, R., Li, Y., Cheng, X., Lin, X., Han, Y.: Natural language understanding and inference with mllm in visual question answering: A survey. ACM Computing Surveys **57**(8), 1–36 (2025)
28. Li, F., Zhang, R., Zhang, H., Zhang, Y., Li, B., Li, W., Ma, Z., Li, C.: Llava-interleave: Tackling multi-image, video, and 3d in large multimodal models. In: ICLR (2024)
29. Li, J., Padmakumar, A., Sukhatme, G., Bansal, M.: Vln-video: Utilizing driving videos for outdoor vision-and-language navigation. In: AAAI. vol. 38, pp. 18517–18526 (2024)
30. Li, Y., Huang, H., Chen, C., Huang, K., Huang, C., Guo, Z., Liu, Z., Xu, J., Li, Y., Li, R., et al.: Migician: Revealing the magic of free-form multi-image grounding in multimodal large language models. In: ACL. pp. 9845–9867 (2025)
31. Liu, F., Emerson, G., Collier, N.: Visual spatial reasoning. TACL **11**, 635–651 (2023)
32. Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Jiang, Q., Li, C., Yang, J., Su, H., et al.: Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In: ECCV. pp. 38–55. Springer (2024)
33. Liu, Y., Duan, H., Zhang, Y., Li, B., Zhang, S., Zhao, W., Yuan, Y., Wang, J., He, C., Liu, Z., et al.: Mmbench: Is your multi-modal model an all-around player? In: ECCV. pp. 216–233. Springer (2024)
34. Majumdar, A., Aggarwal, G., Devnani, B., Hoffman, J., Batra, D.: Zson: Zero-shot object-goal navigation using multimodal goal embeddings. NIPS **35**, 32340–32352 (2022)
35. Malla, S., Choi, C., Dwivedi, I., Choi, J.H., Li, J.: Drama: Joint risk localization and captioning in driving. In: WACV. pp. 1043–1052 (2023)
36. Nguyen, V.D., Do, A.T., Le, D.K., Nguyen, L.V.: Speech recognition-based human–computer interaction: A survey. In: ICTSM. pp. 289–301. Springer (2024)

37. Park, S., Lee, M., Kang, J., Choi, H., Park, Y., Cho, J., Lee, A., Kim, D.: Vlaad: Vision and language assistant for autonomous driving. In: WACV. pp. 980–987 (2024)
38. Patil, V., Sakaridis, C., Liniger, A., Van Gool, L.: P3depth: Monocular depth estimation with a piecewise planarity prior. In: CVPR. pp. 1610–1621 (2022)
39. Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., Rombach, R.: Sdxl: Improving latent diffusion models for high-resolution image synthesis. arXiv preprint arXiv:2307.01952 (2023)
40. Qian, T., Chen, J., Zhuo, L., Jiao, Y., Jiang, Y.G.: Nuscenes-qa: A multi-modal visual question answering benchmark for autonomous driving scenario. In: AAAI. vol. 38, pp. 4542–4550 (2024)
41. Savva, M., Kadian, A., Maksymets, O., Zhao, Y., Wijmans, E., Jain, B., Straub, J., Liu, J., Koltun, V., Malik, J., et al.: Habitat: A platform for embodied ai research. In: ICCV. pp. 9339–9347 (2019)
42. Seaborn, K., Miyake, N.P., Pennefather, P., Otake-Matsuura, M.: Voice in human–agent interaction: A survey. ACM Computing Surveys **54**(4), 1–43 (2021)
43. Shiri, F., Guo, X.Y., Far, M.G., Yu, X., Haf, R., Li, Y.F.: An empirical analysis on spatial reasoning capabilities of large multimodal models. In: EMNLP. pp. 21440–21455 (2024)
44. Sima, C., Renz, K., Chitta, K., Chen, L., Zhang, H., Xie, C., Beißwenger, J., Luo, P., Geiger, A., Li, H.: Drivelm: Driving with graph visual question answering. In: ECCV. pp. 256–274. Springer (2024)
45. Song, S., Li, X., Li, S., Zhao, S., Yu, J., Ma, J., Mao, X., Zhang, W., Wang, M.: How to bridge the gap between modalities: Survey on multimodal large language model. TKDE **37**(9), 5311–5329 (2025)
46. Stogiannidis, I., McDonagh, S., Tsaftaris, S.A.: Mind the gap: Benchmarking spatial reasoning in vision-language models. arXiv preprint arXiv:2503.19707 (2025)
47. Szymańska, E., Dusmanu, M., Buurlage, J.W., Rad, M., Pollefeys, M.: Space3d-bench: Spatial 3d question answering benchmark. In: ECCV. pp. 68–85. Springer (2024)
48. Tapaswi, M., Zhu, Y., Stiefelhagen, R., Torralba, A., Urtasun, R., Fidler, S.: Movieqa: Understanding stories in movies through question-answering. In: CVPR. pp. 4631–4640 (2016)
49. Team, G., Anil, R., Borgeaud, S., Alayrac, J.B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A.M., Hauth, A., Millican, K., et al.: Gemini: a family of highly capable multimodal models. arXiv preprint arXiv:2312.11805 (2023)
50. Wang, F., Fu, X., Huang, J.Y., Li, Z., Liu, Q., Liu, X., Ma, M.D., Xu, N., Zhou, W., Zhang, K., et al.: Muirbench: A comprehensive benchmark for robust multi-image understanding. arXiv preprint arXiv:2406.09411 (2024)
51. Wang, F.E., Yeh, Y.H., Tsai, Y.H., Chiu, W.C., Sun, M.: Bifuse++: Self-supervised and efficient bi-projection fusion for 360 depth estimation. TPAMI **45**(5), 5448–5460 (2022)
52. Wang, Q., He, J., Pan, Y., Yeo, S.Y., Yang, X., Li, S.: Monosr: Open-vocabulary spatial reasoning from monocular images. arXiv preprint arXiv:2511.19119 (2025)
53. Wang, S., Yu, Z., Jiang, X., Lan, S., Shi, M., Chang, N., Kautz, J., Li, Y., Alvarez, J.M.: Omnidrive: A holistic vision-language dataset for autonomous driving with counterfactual reasoning. In: CVPR. pp. 22442–22452 (2025)
54. Wasi, A.T., Faisal, W., Rahman, A., Anik, M.A., Shahriar, M., Topu, M.M., Meem, S.T., Priti, R.N., Mitu, S.A., Hoque, M.I., et al.: Spatialab: Can vision-language models perform spatial reasoning in the wild? arXiv preprint arXiv:2602.03916 (2026)

55. Wu, C., Li, J., Zhou, J., Lin, J., Gao, K., Yan, K., Yin, S.m., Bai, S., Xu, X., Chen, Y., et al.: Qwen-image technical report. arXiv preprint arXiv:2508.02324 (2025)
56. Wu, D., Han, W., Liu, Y., Wang, T., Xu, C.z., Zhang, X., Shen, J.: Language prompt for autonomous driving. In: AAAI. vol. 39, pp. 8359–8367 (2025)
57. Xiao, J., Shang, X., Yao, A., Chua, T.S.: Next-qa: Next phase of question-answering to explaining temporal actions. In: CVPR. pp. 9777–9786 (2021)
58. Xu, Y., Zhu, L., Yang, Y.: Mc-bench: A benchmark for multi-context visual grounding in the era of mllms. In: ICCV. pp. 17675–17687 (2025)
59. Yadav, K., Majumdar, A., Ramrakhya, R., Yokoyama, N., Baevski, A., Kira, Z., Maksymets, O., Batra, D.: Ovrl-v2: A simple state-of-art baseline for imagenav and objectnav. arXiv pre-print:2303.07798 (2023)
60. Yang, J., Yang, S., Gupta, A.W., Han, R., Fei-Fei, L., Xie, S.: Thinking in space: How multimodal large language models see, remember, and recall spaces. In: CVPR. pp. 10632–10643 (2025)
61. Yu, L., Poirson, P., Yang, S., Berg, A.C., Berg, T.L.: Modeling context in referring expressions. In: ECCV. pp. 69–85. Springer (2016)
62. Yue, X., Ni, Y., Zhang, K., Zheng, T., Liu, R., Zhang, G., Stevens, S., Jiang, D., Ren, W., Sun, Y., et al.: Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In: CVPR. pp. 9556–9567 (2024)
63. Zhang, C., Cui, Z., Zhang, Y., Zeng, B., Pollefeys, M., Liu, S.: Holistic 3d scene understanding from a single image with implicit representation. In: CVPR. pp. 8833–8842 (2021)
64. Zhao, Z., Fu, H., Liang, D., Zhou, X., Zhang, D., Xie, H., Wang, B., Bai, X.: Extending large vision-language model for diverse interactive tasks in autonomous driving. arXiv preprint arXiv:2505.08725 (2025)
65. Zhou, S., Vilesov, A., He, X., Wan, Z., Zhang, S., Nagachandra, A., Chang, D., Chen, D., Wang, X.E., Kadambi, A.: Vlm4d: Towards spatiotemporal awareness in vision language models. In: ICCV. pp. 8600–8612 (2025)
66. Zhu, F., Wang, H., Xie, Y., Gu, J., Ding, T., Yang, J., Jiang, H.: Struct2d: A perception-guided framework for spatial reasoning in mllms. Advances in Neural Information Processing Systems **38**, 135805–135837 (2026)
67. Zhu, F., Liang, X., Zhu, Y., Yu, Q., Chang, X., Liang, X.: Soon: Scenario oriented object navigation with graph-based exploration. In: CVPR. pp. 12689–12699 (2021)
68. Zhu, Y., Groth, O., Bernstein, M., Fei-Fei, L.: Visual7w: Grounded question answering in images. In: CVPR. pp. 4995–5004 (2016)