CVSBench: A Comprehensive Benchmark for Cross-view Spatial Reasoning and Dreaming

Ruixun Liu[1*], Lingyu Zhang[1*], Lanxuan Xue[1*], Kaiyu Li[1⋆], Bowen Fu[1], and Xiangyong Cao[1,2⋆]

[1]Faculty of Electronic and Information Engineering, Xi'an Jiaotong University, China
[2]Ministry of Education Key Laboratory of Intelligent Networks and Network Security, China

Abstract. Humans can effortlessly reason about scenes across different viewpoints, yet it remains unclear whether Vision–Language Models (VLMs) possess similar cross-view spatial abilities. Satellite-street scene pairs, with their complex contexts and extreme viewpoint variations, provide an ideal testbed. Motivated by this, we introduce CVSBench, a large-scale benchmark for evaluating cross-view spatial reasoning through satellite-street pairs. This benchmark supports multiple tasks, including cross-view VQA, cross-view grounding, and viewpoint identification. CVSBench comprises 3,297 cross-view image groups with 9,468 object-level annotations and 40,679 question–answer (QA) pairs, enabling systematic and controlled evaluation of cross-view spatial reasoning. Extensive evaluations reveal that advanced VLMs struggle to maintain object-level and layout consistency under drastic viewpoint changes. To bridge this gap towards human-like spatial cognition, we investigate two categories of approaches: spatially grounded reasoning and the incorporation of cognitive map inputs. Our findings demonstrate that language-only reasoning yields marginal improvements, while incorporating visual spatial imagination via a 3D scene imagination pipeline substantially improves cross-view reasoning. These results highlight the necessity of explicit visual-spatial representations for robust spatial cognition in VLMs. Our data and code are released at `https://huggingface.co/datasets/zlyzlyzly/CVSBench`.

Keywords: Cross-view Spatial Reasoning · Vision–Language Models · Benchmark Dataset

1 Introduction

Vision-Language Models (VLMs) [4, 5, 11, 20, 52] have been extensively studied for spatial reasoning recently [9, 10, 35, 65]. However, with the rapid development of embodied navigation, autonomous driving, and urban scene perception,

⋆ Corresponding author
* Equal contribution.

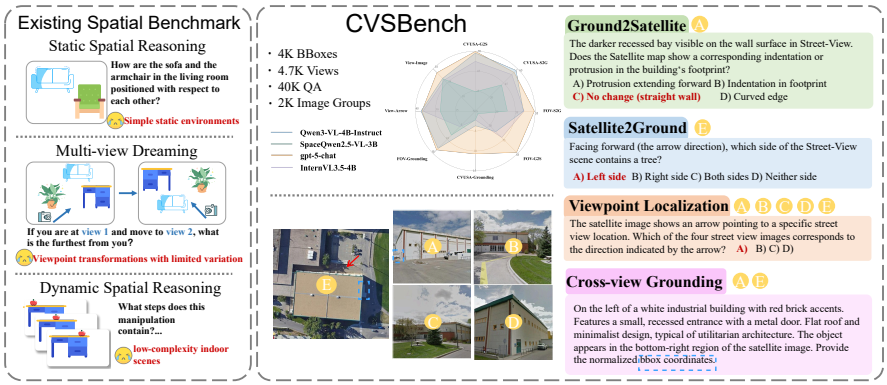

Fig. 1: Overview of **CVSBench**. The left panel reviews representative tasks from existing spatial benchmarks with several limitations [22, 48, 64]. The right panel shows our benchmark specifications, a radar chart of model performance, and core tasks with sample questions, demonstrating CVSBench's superiority in task diversity, scale, and difficulty. The indices (e.g., A–E) after each subtask title indicate the input images. A red arrow indicates viewing perspective, and blue BBoxes denote the target object.

static spatial understanding is no longer sufficient to meet the demands of these tasks [13,21,73]. Humans, for instance, can effortlessly imagine what a street-level scene would look like from an overhead view, or integrate multiple egocentric observations of object distributions into a coherent map-centric layout [6,43]. Even across different coordinate systems or viewing perspectives, humans are able to maintain a consistent internal representation of spatial relationships. Despite the impressive progress achieved by modern VLMs, they still struggle to imagine a scene from an alternative viewpoint given a single observation, as well as to perform reliable cross-view matching of the same objects [16, 64].

To overcome this limitation, recent studies begin to explore the capability of VLMs in understanding dynamic or multi-view scenes [21, 28, 30, 61, 64]. Nevertheless, these efforts still suffer from two major limitations: (1) existing benchmarks are primarily constructed from indoor objects or simple scenes, limiting cross-view evaluation in complex real-world environments [3, 21, 65]; and (2) multi-view settings are often restricted to cameras rotating around an object or the agent itself, resulting in relatively small viewpoint variations [61, 64]. To address these challenges, we adopt satellite–street view scenarios [49, 53, 72], which naturally exhibit complex urban layouts and large viewpoint discrepancies between ground-level and overhead observations. As illustrated in Fig. 1, reasoning in satellite–street view scenarios requires a set of core spatial capabilities, including (i) imagining observations from unseen viewpoints given a single-view input (*e.g.*, whether concave structures on building sides are visible from an overhead view); (ii) maintaining cross-view consistency (*e.g.*, aligning the same objects across views); and (iii) inferring camera viewpoints based on visual cues.

Based on these requirements, we introduce CVSBench, a large-scale benchmark for cross-view spatial reasoning and imagination, comprising the FOV-subset and CVUSA-subset. To obtain accurate annotations and comprehensive evaluation tasks, we design a semi-automatic annotation pipeline that labels cross-view object Bounding Boxes (BBoxes) as well as multi-view alignment relationships. Moreover, we exploit differences in the observation difficulty of the same object across viewpoints to construct QA pairs that probe the spatial imagination of models under limited views. Using this pipeline, CVSBench comprises 3,297 image groups and supports a diverse set of evaluation tasks, including cross-view visual question answering (VQA), grounding, and viewpoint identification. CVSBench enables a systematic evaluation of the spatial cognition in general-purpose VLMs under cross-view scenarios.

Using CVSBench, we conduct extensive evaluations on a wide range of state-of-the-art VLMs and find that existing models still exhibit notable limitations in cross-view scenarios. Inspired by prior work on VLM reasoning and reasoning with visual representations [28, 58, 63, 68], we raise the following question:

Question. *Can VLMs be endowed with stronger cross-view spatial understanding abilities by incorporating explicit reasoning strategies or auxiliary perceptual processes?*

To address this question, we first investigate whether training-free chain-of-thought (CoT) prompting can lead to performance improvements. Building on this analysis, we further propose two CoT-based strategies. (1) Structured Scene CoT, which parses the visual input into a structured textual format. This strategy requires the model to explicitly enumerate object categories and their spatial distributions within the current view before answering the question. (2) Spatial Imagination CoT, which simulates the perspective transformation. This strategy encourages the model to describe the scene composition and layout from the unavailable target viewpoint through textual inference. After training with Supervised Fine-Tuning (SFT) and Reinforcement Learning (RL), the overall gains remain modest, with an average improvement of approximately 1.75% across categories on the FOV-subset, while the improvement on CVUSA-subset is only around 0.4%. Current models struggle to reason about and imagine spatial layouts from alternative viewpoints using textual reasoning alone.

Furthermore, we introduce a 3D scene imagination strategy. Since most existing general-purpose VLMs lack the capability to generate visual intermediate representations [9, 10], we employ depth estimation models and image generation models to synthesize images from 3D viewpoints. These generated images simulate the depth-centered features and "God's-eye-view" cognitive map representations that humans construct mentally. Empirical evaluations demonstrate that, on the FOV-subset, incorporating depth-estimated images leads to a 1.23% improvement, while cognitive map representations yield a 3.34% gain. Our empirical evidence reveals an important finding: compared with purely text-based

reasoning approaches, visual imagination in VLMs has the potential to bring substantially greater improvements to cross-view spatial reasoning tasks.

Our main contributions are summarized as follows:

- We are the first to integrate the **satellite - street view** cross-view VQA task with geo-localization task, enabling a more comprehensive study of the cross-view spatial reasoning and imagination capabilities of VLMs.
- We propose **CVSBench**, a large-scale benchmark dataset featuring extensive human annotations and comprehensive QA pairs, enabling systematic evaluation across multiple cross-view spatial reasoning tasks.
- Through a series of experiments, we demonstrate that the cross-view spatial understanding of current models is primarily constrained by their inability to reason about spatial distributions under viewpoint transformation, and that it can be effectively enhanced through visual spatial imagination strategies.

2 Related Work

2.1 Spatial Thinking evaluation benchmarks

The evaluation for spatial intelligence is undergoing a profound transformation from simple VQA to systematic cognitive assessment, an evolution deeply inspired by theories of spatial thinking components in cognitive psychology [6, 23, 27, 43]. Early benchmarks are primarily confined to 2D relationship judgments [15, 35] or single-modal geometric attribute perception [3, 10, 40]. To address this limitation, the new generation of benchmarks emphasizes cognitive hierarchy and dynamic interaction: recently proposed SIBench [65] and OmniSpatial [22] introduce comprehensive classifications covering perception, understanding, and planning; meanwhile, SPACE [44], systematically examines spatial abilities ranging from navigation to object manipulation. Moreover, VSI-Bench [61] evaluates spatiotemporal memory via video streams, while MINDCUBE [64] further focuses on spatial consistency and mental model construction under multi-view settings. However, existing benchmarks are largely limited to indoor environments, lacking urban-scale evaluations for cross-view localization and logical reasoning between overhead and ground-level perspectives.

2.2 Spatial Understanding and Reasoning in VLMs

While general VLMs have demonstrated remarkable proficiency in semantic understanding and open-ended generation [1, 4, 12, 20, 31, 36, 37, 51], they still suffer from intrinsic deficiencies in precise spatial perception due to the lack of explicit grounding in the physical world, often exhibiting severe "egocentric bias" or "spatial hallucinations" [9, 14, 17, 33, 50]. To enhance spatial reasoning, some studies synthesize large-scale perception data [9] or apply Multimodal CoT mechanisms [41, 69]. Other works inject geometric priors by introducing 3D features [18, 54] or leveraging visual geometry foundation models [56]. Recently, human-like mental simulation has been explored for spatial tasks. For instance,

Fig. 2: Dataset annotation pipeline (Red: ground truth (GT)). **VQA**: Multi-type prompts covering diverse question formats and styles. **Viewpoint Localization**: Red arrows denote camera poses. **Grounding**: Red bboxes, arrows, and dots represent objects, viewpoints, and camera locations.

embodied agents predict physical trajectories [21], and perspective-taking models use mental imagery for cross-view deduction [28]. Additionally, explicit cognitive maps are constructed to help reason under limited viewpoints [64]. However, they currently lack the capacity for "urban-scale" cognition [25,66,71], particularly in processing cross-view data from satellite and street views, which restricts their applicability in broader out-door tasks.

2.3 Remote Sensing VQA and Geo-Grounding tasks

In recent years, multi-modal remote sensing research has achieved remarkable progress in VQA [19, 39] and deep semantic understanding of single-view or cross-view contexts [32, 42, 59]. Building upon these advances, cross-view geo-localization has also evolved from basic feature matching toward complex spatial reasoning that integrates street-view and satellite images [49,53,67,72]. However, current remote sensing visual question answering and cross-view localization remain greatly isolated from each other. To bridge this gap, we present the integration of cross-view localization with VQA, incorporating spatial reasoning to provide a more comprehensive exploration of geospatial intelligence.

3 Benchmark and Evaluation

We introduce **CVSBench**, a benchmark for evaluating spatial reasoning across heterogeneous viewpoints and comprises three task types: *cross-view VQA*, *cross-view grounding*, and *viewpoint selection*.

3.1 Data Annotation Pipeline

CVSBench is constructed from two cross-view datasets: CVUSA [55] and University1652 [70]. We curate 2,155 satellite–panorama image pairs from CVUSA (CVUSA-subset) and extract 1,142 satellite-street view image pairs from University1652 (FOV-subset). A semi-automatic annotation pipeline as shown in Fig. 2 organizes the data into three task families with critical human verification.

Cross-View VQA. The VQA task is structured into two inverse modes: Ground-to-Satellite (G2S) and Satellite-to-Ground (S2G). Structured prompt templates are constructed and combined with paired images to generate candidate questions using *gemini-2.5-flash* [11], covering attributes such as visibility, connectivity, object properties, and spatial relations. Importantly, the two viewpoints provide asymmetric spatial cues: certain attributes are directly observable and easily answered in one view, while they become implicit, partially occluded, or geometrically ambiguous in the other. We leverage this asymmetry to design both the annotation and evaluation protocol. During annotation, the model has access to image pairs to ensure semantic correctness and cross-view consistency. During evaluation, however, it is restricted to a single-view input, requiring the model to infer the missing spatial information through cross-view geometric reasoning rather than direct visual evidence.

Viewpoint Localization. This task is defined only for the FOV-subset. Human annotators mark 2D directional arrows on satellite images, where the start and end points represent camera location and viewing orientation. Based on these geometric priors, two tasks are constructed: (1) View-Arrow, which requires selecting the correct camera pose arrow in a satellite-view image given a street-view image; and (2) View-Image, which requires selecting the correct street-view image given a satellite image with a fixed arrow.

Cross-View Grounding. For the CVUSA-subset, BBox of major objects are first generated in the satellite view using *gemini-2.5-flash* [11]. Based on the object location, a corresponding line-of-sight direction is estimated to search for candidate regions in the ground view, forming an initial cross-view correspondence which is then manually reviewed and corrected. For the FOV-subset, BBoxes are manually annotated. Evaluation includes two formulations: (1) *BBox-to-BBox*: given a BBox in view A, predict the corresponding BBox in view B; (2) *Description-guided grounding*: given a textual description from view A and a coarse region in view B, predict the BBox in view B.

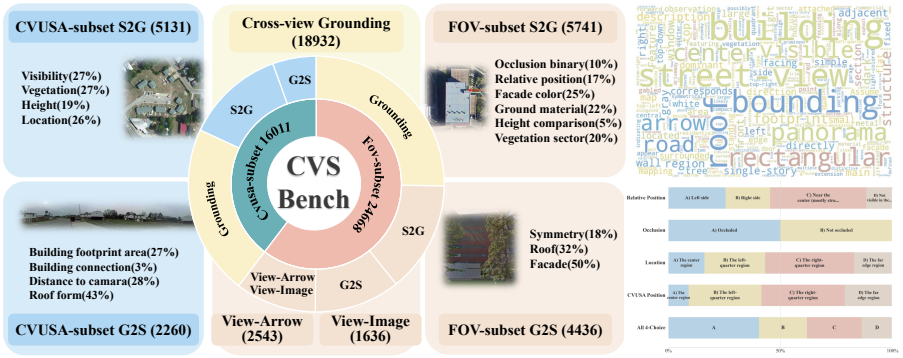

Fig. 3: Overview of task categories in CVSBench. The left panel shows detailed task categorization, with counts or proportions indicated in parentheses. The right panel presents a word cloud of QA composition and the distribution of typical answer categories.

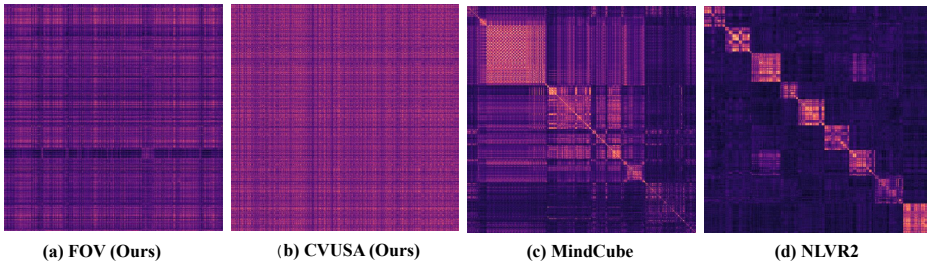

Fig. 4: Similarity heatmaps across datasets. Compared with MindCube and NLVR2, the FOV and CVUSA subsets of CVSBench show lower off-diagonal similarity.

Questions answerable without images are filtered out, and the remaining instances undergo human verification. The review focuses on cross-view identifiability and consistency, while correcting incorrect answers. To mitigate language and answer biases in QA generation, we analyze and balance different option labels. Since evaluation uses single-view settings, the model must reason from limited-viewpoint information, reducing response bias. Eight professional annotators spend approximately 100 hours on annotation and verification, and about 30% of QA pairs are manually corrected. More details are in the Appendix.

To further analyze the diversity of CVSBench, we compare the image similarity heatmaps shown in Fig. 4. MindCube and NLVR2 exhibit more pronounced diagonal and block-wise patterns, indicating that cross-view image pairs within the same group tend to have higher similarity. In contrast, the FOV and CVUSA subsets of CVSBench show lower off-diagonal similarity, suggesting greater cross-view diversity and increased task difficulty.

Table 1: Comparison of CVSBench with existing datasets. The domain categories include RS, General Domain (Gen), and General Spatial (Gen-S). Hyphens (-) denote missing values in the training set. **TI**: Text Instruction, **TA**: Text Answer, **VP**: Visual Prompt. For task capabilities, ✓ indicates support, while ✗ indicates lack thereof.

Category	Benchmark	Train	Test	Anno. Format	Evaluated Capabilities	Query Format	Task Capabilities		
							VQA	Grounding	Multi-view
RS	RSVQA [39]	715K	428K	TI, TA	Basic perception & counting	Short Answer	✓	✗	✗
	RSVGD [67]	23K	15.3K	TI, BBox	Attribute-based grounding	Phrases	✗	✓	✗
	RSIEVAL [19]	-	936	TI, TA	Scene captioning & reasoning	Open-ended	✓	✗	✗
	LHRS-Bench [42]	-	690	TI, TA	Multi-dim semantic understanding	MCQ	✓	✗	✗
	VRSBench [32]	85.8K	37.4K	TI, TA, BBox	OBB grounding & captioning	Open-ended	✓	✓	✗
Gen	MME [14]	-	2.8K	TI, TA	Perception & cognition	T/F	✓	✗	✗
	MMBench [38]	-	3.2K	TI, TA	Fine-grained logical reasoning	MCQ	✓	✗	✗
	SEED-Bench [29]	-	24K	TI, TA	Hierarchical comprehension	MCQ	✓	✗	✓
	VisIT-Bench [7]	-	592	TI, TA	Instruction following	Open-ended	✓	✗	✓
	MC-Bench [60]	-	1.5K	TI, BBox	Multi-image grounding	Referring	✗	✓	✓
Gen-S	VSR [35]	7.6K	3.2K	TI, TA	2D spatial relations	T/F	✓	✗	✗
	BLINK [15]	-	3.9K	TI, TA, VP	Depth & correspondence	MCQ	✓	✓	✓
	SpatialRGPT-Bench [10]	-	1.4K	TI, TA, BBox	Metric & multi-hop logic	Open / MCQ	✓	✓	✗
	VSI-Bench [61]	-	5K	TI, TA	Spatiotemporal memory	MCQ / Num	✓	✗	✓
	3DSRBench [40]	-	2.7K	TI, TA	3D spatial relations	MCQ	✓	✗	✓
	SIBench [65]	-	8.8K	TI, TA	Embodied spatial planning	MCQ / Num	✓	✗	✓
	OmniSpatial [22]	6.9K	1.5K	TI, TA	Dynamic & geometric logic	MCQ / T/F	✓	✗	✓
	NLVR2 [47]	86.4K	7.0K	TI, TA	Compositional visual reasoning	T/F	✓	✗	✓
	MINDCUBE [64]	10K	1.5K	TI, TA	Spatial mental modeling	MCQ	✓	✗	✓
Ours	**CVSBench**	**19.1K**	**21.2K**	**TI, TA, BBox**	**Spatial reasoning & grounding**	**MCQ / T/F / BBox**	✓	✓	✓

3.2 CVSBench Benchmark

Building on the annotation pipeline in Section 3.1, we construct CVSBench for evaluating cross-view spatial reasoning under controlled input protocols. As shown in Fig. 3, the benchmark integrates the CVUSA-subset and FOV-subset and contains 3,297 image groups, 9,468 annotated BBoxes, and 40,679 QA pairs, divided into training and test sets in a 1:1 ratio.

Task Composition. CVSBench comprises several subtask types: Ground-to-Satellite (G2S), Satellite-to-Ground (S2G), Cross-View Grounding (grounding), and Viewpoint Localization (View-Arrow, View-Image). G2S and S2G follow a single-view input protocol, requiring the model to infer spatial attributes of the unseen view. Both four-choice and binary-choice questions are included. Grounding and viewpoint localization assess cross-view entity alignment and camera pose reasoning.

Compared with existing spatial reasoning benchmarks shown in Tab. 1, CVSBench differs in both task design and capability coverage. RS benchmarks focus on object detection or attribute recognition without cross-view correspondence. General Domain (Gen) benchmarks emphasize language and logical reasoning but lack geometric alignment supervision. General spatial (Gen-S) datasets primarily focus on spatial reasoning within a single image (*e.g.*, 2D/3D geometric relations and depth ordering), or only address basic viewpoint transformation tasks where the differences across viewpoints are relatively small. In contrast, CVSBench unifies cross-view VQA, grounding, and viewpoint localization within a single framework. It provides explicit bounding-box supervision, manually annotated camera pose priors, and multi-view symmetry evaluation. As shown in Tab. 1, CVSBench is among the few benchmarks that simultaneously support VQA, grounding, and multi-view settings (cross-view localization), and also demonstrates strong competitiveness in terms of dataset scale.

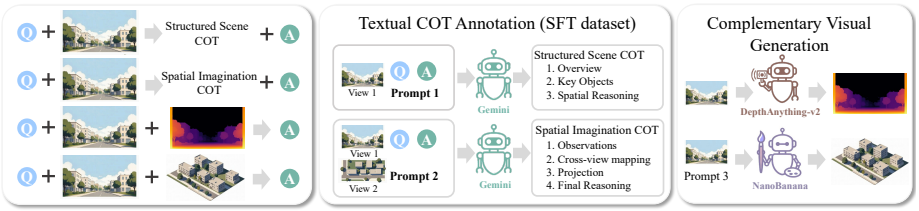

Fig. 5: Framework for enhancing spatial reasoning in VLMs. The center shows the SFT annotation pipeline for textual CoT reasoning, while the right illustrates complementary visual information generation for explicit spatial imagination.

3.3 Improving Visual Spatial Understanding Abilities

Textual Chain-of-Thought Reasoning Given a pre-trained multimodal large language model (MLLM), most prior work adopts a two-stage training paradigm that first performs SFT on annotated CoT data, followed by RL–based optimization. Following this paradigm, we explore two complementary textual CoT strategies with the goal of enhancing the model's spatial reasoning capability and conduct dedicated data annotation, SFT, and RL training for each of them.

Structured CoT. As shown in Fig. 5, we begin by converting visual scenes into structured textual descriptions to facilitate the extraction of fine-grained spatial information. In the process of reasoning, the model is encouraged to explicitly identify and localize objects that are relevant to the final answer. Specifically, the model first outputs the object categories and their corresponding BBoxes, and then performs a lightweight reasoning process over the structured scene representation to derive the answer. To construct the SFT dataset, we provide the input image and the corresponding QA pair to Gemini [11], and prompt it to generate step-by-step structured CoT annotations that include object enumeration, spatial localization, and intermediate reasoning.

Imagination CoT. To better align with the nature of cross-view spatial understanding, we further propose a cross-view imagination CoT strategy. This approach encourages the model to mentally project object layouts observed from the input viewpoint into an alternative viewpoint and perform reasoning accordingly. During SFT data annotation, Gemini2.5-flash [11] is provided with images from both viewpoints together with the QA pair, and is instructed to generate CoT annotations based on the ground-truth object distribution in the target viewpoint. This supervision enables the model to learn how spatial configurations transform across viewpoints through textual reasoning.

Training with GRPO. Based on the annotated datasets, we employ standard SFT followed by RL using the Group Relative Policy Optimization (GRPO) [45] framework. The GRPO objective is defined as:

$$\mathcal{J}_{\text{GRPO}}(\theta) = \mathbb{E}\Big[q \sim P_{\text{sft}}(Q), \{o_i\}_{i=1}^{G} \sim \pi_{\theta_{\text{old}}}(O \mid q)\Big] \\ \frac{1}{G}\sum_{i=1}^{G}\frac{1}{|o_i|}\sum_{t=1}^{|o_i|}\left[\hat{A}^{*}_{i,t} - \gamma\,\mathbb{D}_{\text{KL}}(\pi_\theta \,\|\, \pi_{\theta_{\text{ref}}})\right], \tag{1}$$

where G denotes the number of sampled output sequences, $\hat{A}^{*}_{i,t}$ represents the normalized advantage at token t, and γ controls the strength of KL regularization. The KL divergence term is computed as:

$$\mathbb{D}_{\text{KL}}[\pi_\theta \,\|\, \pi_{\theta_{\text{ref}}}] = \frac{\pi_{\theta_{\text{ref}}}(o_{i,t} \mid q, o_{i,<t})}{\pi_\theta(o_{i,t} \mid q, o_{i,<t})} - \log\frac{\pi_{\theta_{\text{ref}}}(o_{i,t} \mid q, o_{i,<t})}{\pi_\theta(o_{i,t} \mid q, o_{i,<t})} - 1. \tag{2}$$

The reward function is computed as a weighted combination of option-matching accuracy (0.9) and output format compliance (0.1), thereby encouraging the model to produce accurate predictions through reasoning. Additional implementation details and reward specifications are provided in the appendix.

Explicit Spatial Imagination Purely textual descriptions and reasoning are often insufficient to fully represent complex visual scenes. When reasoning about cross-view scenarios, humans can typically construct an "imagined world" directly in the mind and answer questions without relying on explicit textual reasoning [6, 27, 43]. To better simulate this human imagination process, we introduce an image generation model to construct an explicit 3D scene representation analogous to human mental imagery. We adopt nanobanana [11] which is capable of understanding complex instruction to generate a 3D-view image conditioned on the input image. The generated image integrates both side-view information and top-down perspectives, thereby compensating for the inherent incompleteness of visual cues in single-view inputs. In addition, we use a depth estimation model [62] to generate depth maps as the additional input, in order to investigate whether introducing depth information alone, rather than cross-view visual information, can yield similar performance gains. Further details are provided in the Appendix.

4 Experiment

4.1 Experimental Setup

Evaluation Metrics. For multiple-choice tasks, we report accuracy as the proportion of correct answer, while for grounding tasks, we evaluate localization performance using mean Intersection over Union (mIoU) between predicted and ground-truth BBoxes; all results are reported as percentages.

Table 2: Overall comparison on CVSBench. "Acc" means the overall accuracy (%) for VQA and viewpoint tasks, and mIoU (%) for cross-view grounding.

	CVUSA-subset			FOV-subset					Overall	
Method	G2S	S2G	Grounding	G2S	S2G	Grounding	View-Arrow	View-Image	Acc	mIoU
human	88.70	87.53	92.30	85.68	87.20	93.33	98.78	98.34	89.52	93.69
Close-source Models										
gpt-5-chat [46]	69.70	59.50	10.72	53.50	54.10	13.65	40.10	31.70	53.76	12.30
gpt-4o [20]	51.80	62.10	10.80	36.00	46.80	13.61	30.00	26.60	45.54	12.30
Open-source Models										
Deepseekv3.2 [34]	63.39	47.03	3.29	56.69	38.68	10.98	26.13	23.92	44.43	7.39
LLaVA-OV-1.5-4B-Instruct [2]	68.98	60.09	9.61	21.32	41.19	8.56	27.95	24.22	41.51	9.06
Gemma-3-4b-it [24]	71.00	52.88	0.38	40.51	47.99	6.94	30.13	28.68	46.43	3.87
InternVL3.5-4B [52]	58.10	56.00	2.22	34.30	44.80	5.53	39.30	25.70	44.28	3.99
InternVL3.5-8B [52]	65.20	61.90	2.59	52.30	41.50	4.69	40.10	26.80	49.89	3.71
Qwen3-VL-8B-Instruct [4]	69.90	61.90	0.60	24.00	46.70	8.10	42.30	27.50	45.88	4.59
Qwen3-VL-4B-Instruct [4]	67.90	62.40	3.68	26.60	44.20	13.15	40.80	27.10	45.48	8.72
Qwen2.5-VL-3B [5]	44.90	55.20	6.18	19.00	35.80	13.46	30.60	23.60	36.48	10.06
Geochat-7B [26]	60.69	56.12	0.30	56.19	44.44	0.25	25.59	28.68	48.23	0.27
Spatial Reasoning Models										
SpaceQwen2.5VL-3B [9]	22.70	40.20	2.22	3.40	29.40	8.50	26.40	26.70	25.45	5.56
SpaceThinker-Qwen2.5VL-3B [9]	21.80	34.10	5.46	19.60	25.30	8.45	32.40	23.40	26.66	7.05
SpatialBot-3B [8]	68.30	59.08	0.00	29.18	42.38	0.02	26.50	28.68	43.35	0.01
ViLaSR-7B [57]	68.20	85.20	1.28	35.14	41.79	12.01	25.41	27.49	44.05	6.39

Training Protocol. To analyze reasoning mechanisms, we use Qwen3-VL-4B [4] as our base model for controlled training. The training follows two stages: SFT and RL. Half of the training data is used for SFT with Structured Scene CoT or Spatial Imagination CoT. The temperature for sampling is set to 0.01. The left portion of the dataset is used for RL training, with the temperature fixed at 0.7. Learning rate is set to 3e-5 in the SFT stage and 1e-6 in the RL stage.

4.2 Main Benchmark Results

Human Baseline. To calibrate task difficulty, we collect human performance on the test set. The questions are presented through a unified web-based UI, and eight persons independently provide answers. We report the aggregated accuracy across annotators. As shown in Tab. 2, human performance substantially exceeds current VLMs, indicating that the benchmark remains solvable while being challenging for existing models.

Model-Level Comparison. Tab. 2 presents the overall performance on CVSBench. Closed-source models achieve the strongest results, with gpt-5-chat ranking highest. Among open-source systems, InternVL3.5-8B [52] performs best and approaches closed-source performance on VQA tasks. In contrast, spatial reasoning–oriented models (*e.g.* , SpaceQwen [9] and SpaceThinker [9]) underperform general-purpose VLMs overall. Although these models are designed to enhance single-view spatial understanding through stronger geometric and depth-aware reasoning, such improvements do not directly translate into better cross-view correspondence. Moreover, their ability to recognize attributes that rely on fine-grained texture details cues such as facades and materials is relatively weak. Consequently, these models remain limited in tasks that require maintaining structural and semantic consistency across multiple viewpoints.

Table 3: Effect of inference-time CoT prompting on Qwen3-VL-4B [4]. The model is required to first output reasoning traces before producing the final answer.

	CVUSA-subset			FOV-subset				
CoT	G2S	S2G	Grounding	G2S	S2G	Grounding	View-Arrow	View-Image
None	67.90	62.40	3.68	26.70	43.90	13.15	40.80	27.10
Inference CoT	69.20	61.70	6.01	35.60	45.30	12.39	21.90	23.60

Table 4: Performance comparison (%) of different training strategies and CoT designs on G2S and S2G of CVUSA-subset. Top two results are highlighted in green: dark green for the first and light green for the second. Footprint: Building footprint area; Connection: Building connection; Distance: Distance to camera; Roof: Roof form.

		CVUSA-subset G2S				CVUSA-subset S2G			
Training	CoT	Footprint	Connection	Distance	Roof	Height	Location	Vegetation	Visibility
None	None	62.8	25.0	84.0	63.2	59.5	64.0	77.4	47.9
SFT	Structured Scene	63.2	28.1	88.7	50.5	53.3	63.4	74.8	57.5
+ RL	Structured Scene	58.6	37.2	87.5	49.2	53.3	65.1	73.1	55.9
SFT	Spatial Imagination	64.6	31.3	89.1	46.5	50.3	62.0	73.9	52.4
+ RL	Spatial Imagination	63.2	31.2	89.8	62.0	48.7	64.6	74.6	53.3

Task-Level Difficulty. Cross-view VQA tasks (G2S and S2G) exhibit moderate performance, suggesting that coarse-grained cross-view associations are partially solvable. In contrast, grounding tasks consistently achieve low IoU scores, indicating insufficient capability in establishing cross-view object correspondence. This limitation can be attributed to two factors: the lack of training on multi-image grounding scenarios and the substantial appearance variation of the same object across different viewpoints. As a result, models struggle to achieve precise entity-level alignment across views. Furthermore, performance under the FOV-subset is consistently lower than that on CVUSA-subset, demonstrating that a restricted field of view and incomplete structural cues further exacerbate the difficulty of cross-view correspondence.

View Alignment Phenomenon. In the Viewpoint Localization task, arrow selection is consistently easier than image selection. The former involves matching multiple arrows in a RS image with a single street-view image, whereas the latter requires modeling the correspondence between four candidate street-view images and a single arrow in the RS image. Consequently, image selection demands processing a larger amount of visual information to extract feature correspondences. This performance gap highlights the limitations of VLMs in handling cross-view visual feature representation and entity-level alignment.

4.3 Textual Reasoning Exploration

We investigate whether modifying textual reasoning improves cross-view spatial understanding. We prompt the model to reason at inference time without further

Table 5: Performance comparison (%) of different training strategies and CoT designs on G2S and S2G of FOV-subset. Color: Facade color; Material: Ground material; Height: Height comparison; Occlusion: Occlusion binary; Position: Relative position.

		FOV-subset G2S			FOV-subset S2G					
Training	CoT	Facade	Roof	Symmetry	Color	Material	Height	Occlusion	Position	Vegetation Sector
None	None	18.4	9.9	79.8	24.5	41.5	73.6	51.9	56.1	49.0
SFT	Structured Scene	41.3	14.8	72.9	32.0	46.2	65.5	48.1	56.1	51.6
+ RL	Structured Scene	25.5	11.2	56.0	29.0	45.4	64.2	42.4	58.7	51.7
SFT	Spatial Imagination	51.0	14.1	57.1	27.7	43.8	66.7	54.8	55.8	49.5
+ RL	Spatial Imagination	43.3	8.4	67.5	27.0	43.9	69.6	42.4	55.9	51.0

Table 6: Performance comparison (%) of different auxiliary views for Qwen3VL-4B on the FOV-subset G2S and S2G tasks.

	FOV-subset G2S			FOV-subset S2G					
Auxiliary View	Facade	Roof	Symmetry	Color	Material	Height	Occlusion	Position	Vegetation Sector
None	18.4	9.9	79.8	24.5	41.5	73.6	51.9	56.1	49.0
Depth	18.5	10.2	79.8	24.4	42.2	74.3	54.8	56.1	48.8
3D View	17.1	18.7	80.5	36.9	47.0	75.0	51.2	58.3	50.0

training. As shown in Tab. 3, inference-time CoT yields marginal gains in certain categories, particularly grounding and FOV-subset G2S. Grounding under CVUSA shows slight improvement, suggesting that reasoning can partially guide spatial correspondence. However, improvements are inconsistent across different tasks, indicating that expanding reasoning during inference alone cannot fundamentally improve cross-view alignment. We further train and evaluate models under two supervised reasoning formats: Cross-View Imagination CoT and Structured Scene CoT, where each format is optimized through SFT followed by RL. These augmentation strategies are applied only to cross-view VQA tasks, and not to grounding or viewpoint localization.

Structured CoT. Structured CoT enforces object-centric decomposition by grounding entities before attribute reasoning. As shown in Tab. 4 and Tab. 5, it improves several attributes that require cross-view entity correspondence and spatial alignment. These persistent improvements suggest that grounding-first decomposition provides a relatively stable reasoning scaffold rather than transient gains driven by superficial cues. Fig. 6 shows representative cases. For example, it separates roof regions from facade regions when predicting facade color from satellite views, and distinguishes perspective effects when comparing building footprints across views. By disentangling regions before attribute prediction, it mitigates cross-view attribute confusion and leads to more stable improvements.

Imagination CoT. This paradigm introduces explicit projection from ground view to a hypothesized satellite representation. On CVUSA-subset, it improves geometry-sensitive categories such as distance estimation, suggesting better approximation of latent 3D correspondence. As shown in Fig. 6, imagination-style projection can correctly reason about footprint span, whereas Structured Scene CoT remains more susceptible to perspective-induced bias. However, gains on

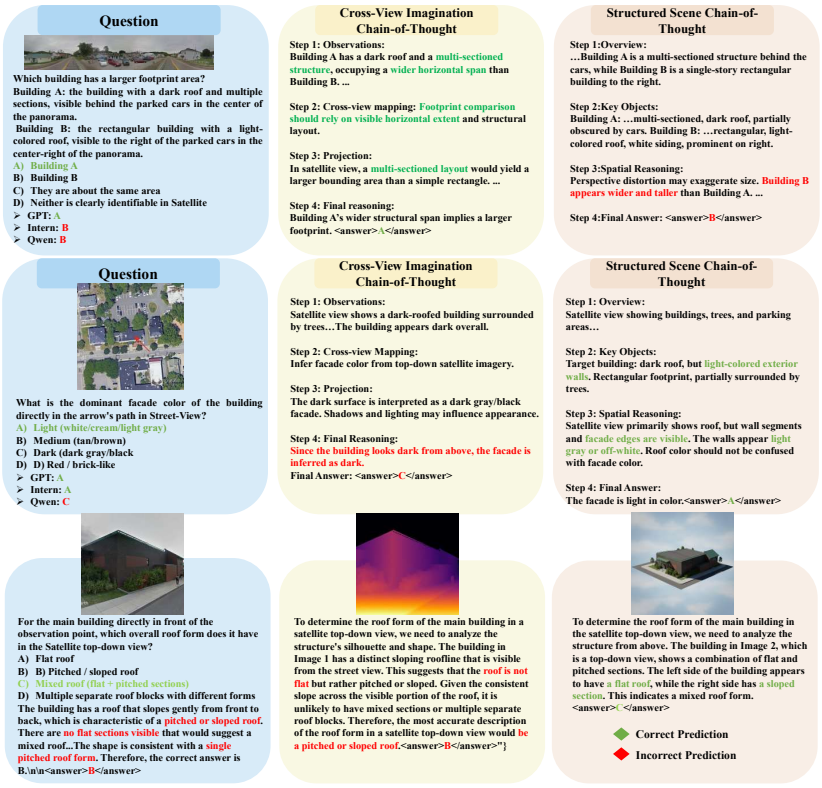

Fig. 6: Qualitative examples. Comparison of methods for enhancing VLM spatial understanding. Green indicates GT or correct prediction. Red indicates incorrect prediction.

appearance-dominant attributes are less consistent across datasets and remain limited on FOV-subset.

Additionally, on the FOV-subset, performance on certain types drops after RL compared with SFT. These tasks rely on fine-grained visual cues that are hard to resolve through textual reasoning alone (*e.g.*, facade color, roof type, and occlusion depend on boundaries, inclination, and viewpoint projection). As a result, the model struggles to learn reliable reasoning strategies and develops preferences for specific answer options during evaluation, leading to reward hacking.

4.4 Explicit Spatial Representation Exploration

We investigate whether stronger spatial representations at the input level lead to more stable improvements. We apply 3D-miniature and depth augmentations only to the FOV-subset because it provides limited visual cues and benefits more

from additional 3D information, whereas the CVUSA-subset already offers richer panoramic context and reliable 3D-view generation from panoramic imagery remains difficult for existing image generation models.

Depth Augmentation. Depth augmentation primarily enriches local geometric ordering within the same viewpoint. As shown in Tab. 6, by making foreground-background relations explicit, it mainly improves geometry-sensitive categories such as occlusion and height in the FOV-subset. However, gains on appearance-dependent attributes (e.g., facade and material) remain limited, suggesting that single-view depth cues refine local geometry but do not substantially improve cross-view structural alignment.

3D Miniature Rendering. In contrast, the 3D miniature view introduces a more explicit global structural representation. As shown in Tab. 6, it yields broader gains in facade color recognition, position estimation, and roof-related categories. Rather than refining local geometry, it provides a compact structural abstraction closer to satellite perspective, indicating that modeling global structural layout is more effective for cross-view correspondence than strengthening single-view geometric detail alone. As shown in Fig. 6, single-view cues may mislead global roof inference. Depth improves local geometry, whereas the 3D view reveals global structure, explaining its stronger gains over depth and text-only CoT. More details are in the Appendix.

5 Conclusion

We present CVSBench, a large-scale benchmark for evaluating cross-view spatial reasoning and imagination in VLMs under the satellite–street view setting. Our study shows that current VLMs struggle to maintain consistent spatial representations under large viewpoint changes, revealing a limitation beyond static spatial understanding. Across several reasoning strategies, text-based approaches offer limited gains, while visual spatial imagination shows greater potential for enhancing cross-view reasoning. These findings suggest that future VLMs should move beyond language-only reasoning and incorporate explicit perceptual imagination mechanisms. We hope CVSBench will facilitate further research toward more robust and human-like spatial cognition in VLMs.

Acknowledgements

This work is supported by the Fundamental and Interdisciplinary Disciplines Breakthrough Plan of the Ministry of Education of China (No. JYB2025XDXM101), China NSFC Projects under Contract 62272375, and the Tianyuan Fund for Mathematics of the National Natural Science Foundation of China (Grant No. 12426105).

References

1. Alayrac, J.B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., et al.: Flamingo: a visual language model for few-shot learning. Advances in neural information processing systems **35**, 23716–23736 (2022)
2. An, X., Xie, Y., Yang, K., et al.: Llava-onevision-1.5: Fully open framework for democratized multimodal training. In: arxiv (2025)
3. Azuma, D., Miyanishi, T., Kurita, S., Kawanabe, M.: Scanqa: 3d question answering for spatial scene understanding. In: proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 19129–19139 (2022)
4. 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)
5. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., Zhong, H., Zhu, Y., et al.: Qwen2.5-vl technical report. arXiv preprint arXiv:2502.13923 (2025)
6. Bednarz, R., Lee, J.: What improves spatial thinking? evidence from the spatial thinking abilities test. International Research in Geographical and environmental education **28**(4), 262–280 (2019)
7. Bitton, Y., Bansal, H., Hessel, J., Shao, R., Zhu, W., Awadalla, A., Gardner, J., Taori, R., Schmidt, L.: Visit-bench: A benchmark for vision-language instruction following inspired by real-world use. arXiv preprint arXiv:2308.06595 (2023)
8. Cai, W., Ponomarenko, I., Yuan, J., Li, X., Yang, W., Dong, H., Zhao, B.: Spatialbot: Precise spatial understanding with vision language models. In: 2025 IEEE International Conference on Robotics and Automation (ICRA). pp. 9490–9498. IEEE (2025)
9. Chen, B., Xu, Z., Kirmani, S., Ichter, B., Sadigh, D., Guibas, L., Xia, F.: Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14455–14465 (2024)
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. Advances in Neural Information Processing Systems **37**, 135062–135093 (2024)
11. Comanici, G., Bieber, E., Schaekermann, M., Pasupat, I., Sachdeva, N., Dhillon, I., Blistein, M., Ram, O., Zhang, D., Rosen, E., et al.: Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. arXiv preprint arXiv:2507.06261 (2025)
12. Dai, W., Li, J., Li, D., Tiong, A., Zhao, J., Wang, W., Li, B., Fung, P.N., Hoi, S.: Instructblip: Towards general-purpose vision-language models with instruction tuning. Advances in neural information processing systems **36**, 49250–49267 (2023)
13. Ding, X., Han, J., Xu, H., Liang, X., Zhang, W., Li, X.: Holistic autonomous driving understanding by bird's-eye-view injected multi-modal large models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13668–13677 (2024)
14. Fu, C., Chen, P., Shen, Y., Qin, Y., Zhang, M., Lin, X., Yang, J., Zheng, X., Li, K., Sun, X., et al.: Mme: A comprehensive evaluation benchmark for multimodal large language models. arXiv preprint arXiv:2306.13394 (2023)
15. 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: European Conference on Computer Vision. pp. 148–166. Springer (2024)

16. Gholami, M., Rezaei, A., Weimin, Z., Mao, S., Zhou, S., Zhang, Y., Akbari, M.: Spatial reasoning with vision-language models in ego-centric multi-view scenes. arXiv preprint arXiv:2509.06266 (2025)
17. Guan, T., Liu, F., Wu, X., Xian, R., Li, Z., Liu, X., Wang, X., Chen, L., Huang, F., Yacoob, Y., et al.: Hallusionbench: an advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 14375–14385 (2024)
18. Hong, Y., Zhen, H., Chen, P., Zheng, S., Du, Y., Chen, Z., Gan, C.: 3d-llm: Injecting the 3d world into large language models. Advances in Neural Information Processing Systems **36**, 20482–20494 (2023)
19. Hu, Y., Yuan, J., Wen, C., Lu, X., Liu, Y., Li, X.: Rsgpt: A remote sensing vision language model and benchmark. ISPRS Journal of Photogrammetry and Remote Sensing **224**, 272–286 (2025)
20. Hurst, A., Lerer, A., Goucher, A.P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A., Welihinda, A., Hayes, A., Radford, A., et al.: Gpt-4o system card. arXiv preprint arXiv:2410.21276 (2024)
21. Ji, Y., Tan, H., Shi, J., Hao, X., Zhang, Y., Zhang, H., Wang, P., Zhao, M., Mu, Y., An, P., et al.: Robobrain: A unified brain model for robotic manipulation from abstract to concrete. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1724–1734 (2025)
22. 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)
23. Johnson-Laird, P.N.: Mental models in cognitive science. Cognitive science **4**(1), 71–115 (1980)
24. Khandoga, M., Kostiuk, Y., Polishko, A., Kozlov, K., Filipchuk, Y., Kiulian, A.: Framing the language: Fine-tuning gemma 3 for manipulation detection pp. 49–54 (2025)
25. Kim, M.J., Pertsch, K., Karamcheti, S., Xiao, T., Balakrishna, A., Nair, S., Rafailov, R., Foster, E., Lam, G., Sanketi, P., et al.: Openvla: An open-source vision-language-action model. arXiv preprint arXiv:2406.09246 (2024)
26. Kuckreja, K., Danish, M.S., Naseer, M., Das, A., Khan, S., Khan, F.S.: Geochat: Grounded large vision-language model for remote sensing (2023)
27. Lee, J., Bednarz, R.: Components of spatial thinking: Evidence from a spatial thinking ability test. Journal of geography **111**(1), 15–26 (2012)
28. Lee, P.Y., Je, J., Park, C., Uy, M.A., Guibas, L., Sung, M.: Perspective-aware reasoning in vision-language models via mental imagery simulation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9241–9251 (2025)
29. Li, B., Ge, Y., Ge, Y., Wang, G., Wang, R., Zhang, R., Shan, Y.: Seed-bench: Benchmarking multimodal large language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13299–13308 (2024)
30. Li, D., Li, H., Wang, Z., Yan, Y., Zhang, H., Chen, S., Hou, G., Jiang, S., Zhang, W., Shen, Y., et al.: Viewspatial-bench: Evaluating multi-perspective spatial localization in vision-language models. arXiv preprint arXiv:2505.21500 (2025)
31. Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In: International conference on machine learning. pp. 19730–19742. PMLR (2023)

32. Li, X., Ding, J., Elhoseiny, M.: Vrsbench: A versatile vision-language benchmark dataset for remote sensing image understanding. Advances in Neural Information Processing Systems **37**, 3229–3242 (2024)
33. Li, Y., Du, Y., Zhou, K., Wang, J., Zhao, W.X., Wen, J.R.: Evaluating object hallucination in large vision-language models. In: Proceedings of the 2023 conference on empirical methods in natural language processing. pp. 292–305 (2023)
34. Liu, A., Mei, A., Lin, B., Xue, B., Wang, B., Xu, B., Wu, B., Zhang, B., Lin, C., Dong, C., et al.: Deepseek-v3. 2: Pushing the frontier of open large language models (2025)
35. Liu, F., Emerson, G., Collier, N.: Visual spatial reasoning. Transactions of the Association for Computational Linguistics **11**, 635–651 (2023)
36. Liu, H., Li, C., Li, Y., Lee, Y.J.: Improved baselines with visual instruction tuning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 26296–26306 (2024)
37. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. Advances in neural information processing systems **36**, 34892–34916 (2023)
38. 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: European conference on computer vision. pp. 216–233. Springer (2024)
39. Lobry, S., Marcos, D., Murray, J., Tuia, D.: Rsvqa: Visual question answering for remote sensing data. IEEE Transactions on Geoscience and Remote Sensing **58**(12), 8555–8566 (2020)
40. Ma, W., Chen, H., Zhang, G., Chou, Y.C., Chen, J., de Melo, C., Yuille, A.: 3dsrbench: A comprehensive 3d spatial reasoning benchmark. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6924–6934 (2025)
41. Mitra, C., Huang, B., Darrell, T., Herzig, R.: Compositional chain-of-thought prompting for large multimodal models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14420–14431 (2024)
42. Muhtar, D., Li, Z., Gu, F., Zhang, X., Xiao, P.: Lhrs-bot: Empowering remote sensing with vgi-enhanced large multimodal language model. In: European Conference on Computer Vision. pp. 440–457. Springer (2024)
43. Newcombe, N.S., Shipley, T.F.: Thinking about spatial thinking: New typology, new assessments. In: Studying visual and spatial reasoning for design creativity, pp. 179–192. Springer (2014)
44. Ramakrishnan, S.K., Wijmans, E., Kraehenbuehl, P., Koltun, V.: Does spatial cognition emerge in frontier models? arXiv preprint arXiv:2410.06468 (2024)
45. Shao, Z., Wang, P., Zhu, Q., Xu, R., Song, J., Bi, X., Zhang, H., Zhang, M., Li, Y., Wu, Y., et al.: Deepseekmath: Pushing the limits of mathematical reasoning in open language models. arXiv preprint arXiv:2402.03300 (2024)
46. Singh, A., Fry, A., Perelman, A., Tart, A., Ganesh, A., El-Kishky, A., McLaughlin, A., Low, A., Ostrow, A., Ananthram, A., et al.: Openai gpt-5 system card. arXiv preprint arXiv:2601.03267 (2025)
47. Suhr, A., Zhou, S., Zhang, A., Zhang, I., Bai, H., Artzi, Y.: A corpus for reasoning about natural language grounded in photographs. In: Proceedings of the 57th annual meeting of the association for computational linguistics. pp. 6418–6428 (2019)
48. Szymańska, E., Dusmanu, M., Buurlage, J.W., Rad, M., Pollefeys, M.: Space3dbench: Spatial 3d question answering benchmark. In: European Conference on Computer Vision. pp. 68–85. Springer (2024)
49. Toker, A., Zhou, Q., Maximov, M., Leal-Taixé, L.: Coming down to earth: Satellite-to-street view synthesis for geo-localization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6488–6497 (2021)

50. Tong, S., Liu, Z., Zhai, Y., Ma, Y., LeCun, Y., Xie, S.: Eyes wide shut? exploring the visual shortcomings of multimodal llms. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9568–9578 (2024)
51. Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., Chen, K., Liu, X., Wang, J., Ge, W., et al.: Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. arXiv preprint arXiv:2409.12191 (2024)
52. Wang, W., Gao, Z., Gu, L., Pu, H., Cui, L., Wei, X., Liu, Z., Jing, L., Ye, S., Shao, J., et al.: Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. arXiv preprint arXiv:2508.18265 (2025)
53. Wang, Y., Liu, Z., Wang, Z., Hu, H., Liu, P., Rao, Y.: Geovista: Web-augmented agentic visual reasoning for geolocalization. arXiv preprint arXiv:2511.15705 (2025)
54. Wang, Z., Huang, H., Zhao, Y., Zhang, Z., Zhao, Z.: Chat-3d: Data-efficiently tuning large language model for universal dialogue of 3d scenes. arXiv preprint arXiv:2308.08769 (2023)
55. Workman, S., Souvenir, R., Jacobs, N.: Wide-area image geolocalization with aerial reference imagery. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 3961–3969 (2015)
56. Wu, D., Liu, F., Hung, Y.H., Duan, Y.: Spatial-mllm: Boosting mllm capabilities in visual-based spatial intelligence. arXiv preprint arXiv:2505.23747 (2025)
57. Wu, J., Guan, J., Feng, K., Liu, Q., Wu, S., Wang, L., Wu, W., Tan, T.: Reinforcing spatial reasoning in vision-language models with interwoven thinking and visual drawing (2025)
58. Wu, W., Mao, S., Zhang, Y., Xia, Y., Dong, L., Cui, L., Wei, F.: Mind's eye of llms: visualization-of-thought elicits spatial reasoning in large language models. Advances in Neural Information Processing Systems **37**, 90277–90317 (2024)
59. Xu, H., Hu, Y., Zhu, Z., Gao, C., Wang, Z., Rao, J., Lu, W., Li, W., Yin, Q., Li, Y.: Citycube: Benchmarking cross-view spatial reasoning on vision-language models in urban environments. arXiv preprint arXiv:2601.14339 (2026)
60. Xu, Y., Zhu, L., Yang, Y.: Mc-bench: A benchmark for multi-context visual grounding in the era of mllms. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 17675–17687 (2025)
61. 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: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 10632–10643 (2025)
62. Yang, L., Kang, B., Huang, Z., Zhao, Z., Xu, X., Feng, J., Zhao, H.: Depth anything v2. Advances in Neural Information Processing Systems **37**, 21875–21911 (2024)
63. Yang, Z., Yu, X., Chen, D., Shen, M., Gan, C.: Machine mental imagery: Empower multimodal reasoning with latent visual tokens. arXiv preprint arXiv:2506.17218 (2025)
64. Yin, B., Wang, Q., Zhang, P., Zhang, J., Wang, K., Wang, Z., Zhang, J., Chandrasegaran, K., Liu, H., Krishna, R., et al.: Spatial mental modeling from limited views. In: Structural Priors for Vision Workshop at ICCV'25 (2025)
65. Yu, S., Chen, Y., Ju, H., Jia, L., Zhang, F., Huang, S., Wu, Y., Cui, R., Ran, B., Zhang, Z., et al.: How far are vlms from visual spatial intelligence? a benchmark-driven perspective. arXiv preprint arXiv:2509.18905 (2025)
66. Yuan, W., Duan, J., Blukis, V., Pumacay, W., Krishna, R., Murali, A., Mousavian, A., Fox, D.: Robopoint: A vision-language model for spatial affordance prediction for robotics. arXiv preprint arXiv:2406.10721 (2024)

67. Zhan, Y., Xiong, Z., Yuan, Y.: Rsvg: Exploring data and models for visual grounding on remote sensing data. IEEE transactions on geoscience and remote sensing **61**, 1–13 (2023)
68. Zhang, L., Zhai, X., Zhao, Z., Zong, Y., Wen, X., Zhao, B.: What if the tv was off? examining counterfactual reasoning abilities of multi-modal language models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 21853–21862 (2024)
69. Zhang, Z., Zhang, A., Li, M., Zhao, H., Karypis, G., Smola, A.: Multimodal chain-of-thought reasoning in language models. arXiv preprint arXiv:2302.00923 (2023)
70. Zheng, Z., Wei, Y., Yang, Y.: University-1652: A multi-view multi-source benchmark for drone-based geo-localization. In: Proceedings of the 28th ACM international conference on Multimedia. pp. 1395–1403 (2020)
71. Zhu, C., Wang, T., Zhang, W., Pang, J., Liu, X.: Llava-3d: A simple yet effective pathway to empowering lmms with 3d-awareness. arXiv preprint arXiv:2409.18125 (2024)
72. Zhu, S., Yang, T., Chen, C.: Vigor: Cross-view image geo-localization beyond one-to-one retrieval. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3640–3649 (2021)
73. Zhu, Z., Wang, X., Li, Y., Zhang, Z., Ma, X., Chen, Y., Jia, B., Liang, W., Yu, Q., Deng, Z., et al.: Move to understand a 3d scene: Bridging visual grounding and exploration for efficient and versatile embodied navigation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8120–8132 (2025)