CityCube: Benchmarking Cross-view Spatial Reasoning on Vision-Language Models in Urban Environments

Anonymous ACL submission

Abstract

Cross-view spatial reasoning is essential for embodied AI, underpinning spatial understanding, mental simulation and planning in complex environments. Existing benchmarks primarily emphasize indoor or street settings, overlooking the unique challenges of open-ended urban spaces characterized by rich semantics, complex geometries, and view variations. To address this, we introduce **CityCube**, a systematic benchmark designed to probe cross-view reasoning capabilities of current VLMs in urban settings. CityCube integrates four viewpoint dynamics to mimic camera movements and spans a wide spectrum of perspectives from multiple platforms, e.g., vehicles, drones and satellites. For a comprehensive assessment, it features 5,022 meticulously annotated multi-view QA pairs categorized into five cognitive dimensions and three spatial relation expressions. A comprehensive evaluation of 33 VLMs reveals a significant performance disparity with humans: even large-scale models struggle to exceed 54.1% accuracy, remaining 34.2% below human performance. By contrast, small-scale fine-tuned VLMs achieve over 60.0% accuracy, highlighting the necessity of our benchmark. Further analyses indicate the task correlations and fundamental cognitive disparity between VLMs and human-like reasoning.

1 Introduction

Spatial reasoning across viewpoints and scales is fundamental to spatial intelligence. It goes beyond geometric measures (Yang et al., 2024; Cai et al., 2025a), also involves abilities such as relational reasoning (Chen et al., 2024; Song et al., 2025), perspective taking (Piaget, 2013; Li et al., 2025), mental simulation (Eslami et al., 2018), dynamic perception (Tversky, 2019; Ding et al., 2025) and world knowledge recalling (Jia et al., 2025). While humans naturally perform tasks like reasoning 3D scenes from streaming 2D views, replicating this in embodied AI remains challenging.

Recently, Vision-language Models (VLMs) have been increasingly adopted as the cognitive backbone for embodied agents, such as drones, mobile robots and autonomous vehicles (Majumdar et al., 2024). These agents dynamically operate and interact with the physical world, naturally requiring VLMs to perceive and understand multiple views (Hong et al., 2023; Zhang et al., 2024; Zhu et al., 2024; Qi et al., 2024). However, whether current VLMs possess the requisite capabilities for expansive urban spaces remains an open question, as comprehensive evaluations in this domain are still lacking.

To address this, we argue that benchmarks must extend beyond existing indoor-focused settings (Yang et al., 2025a; Du et al., 2024) to the broader context of urban open spaces. As shown in Fig. 1(a), urban environments present unique challenges for cross-view spatial intelligence (**CvSI**):

- **Richer semantics:** Dense and repetitive instances (e.g., signage and vehicles) pose strict demands on VLMs in spatial grounding and disambiguation.
- **Complex geometries:** Intricate urban 3D structures and road networks demand robust spatial mental reconstruction capabilities.
- **Cross-scale viewpoint variations:** Diverse perspectives, from egocentric to top-down (e.g., drones), necessitate rigorous cross-scale reasoning to maintain spatial consistency.

To the best of our knowledge, CvSI in urban embodied tasks remains underexplored. To address this gap, we introduce **CityCube** (Fig. 1), a systematical benchmark for cross-view urban spatial reasoning. CityCube is constructed by integrating urban views from real-world datasets and realistic simulators, forming a large-scale dataset with 18K images. It covers more than 70 representative

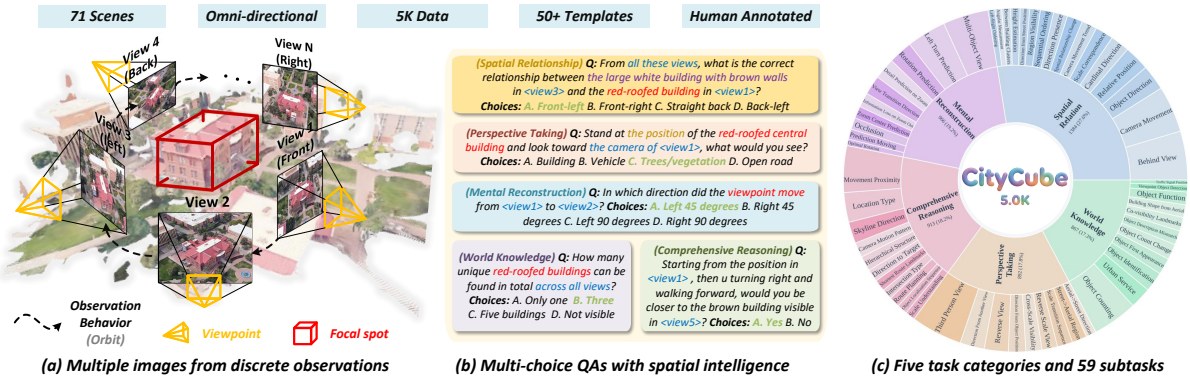


Figure 1: Illustration of the CityCube benchmark. **Left:** An illustration of an embodied orbiting observation, where an agent captures multi-view images by circling a focal object (highlighted in red). **Middle:** Examples of multi-choice QA designed to evaluate five dimensions of CvSI. **Right:** Task distributions on CityCube Benchmark.

cities, and includes multi-view imagery captured from heterogeneous platforms, providing a wide spectrum of first-person and aerial perspectives.

Based on this foundation, we design a comprehensive and challenging suite of spatial reasoning tasks to systematically probe CvSI of urban embodiments. Specifically, CityCube evaluates 5,022 QA pairs with five fundamental spatial intelligences in urban embodied scenes (Cai et al., 2025b), as illustrated in Fig. 1(b): Spatial Relations (**SR**), Perspective Taking (**PT**), Mental Reconstruction (**MR**), World Knowledge (**WK**), and Comprehensive Reasoning (**CR**). Each ability requires cross-scale and multi-centric spatial reasoning over diverse scenes, reflecting the strengths and weaknesses of an agent in geometry, semantics, viewpoint transformation, and contextual knowledge.

Upon these tasks, we conduct a large-scale evaluation of 33 mainstream VLMs. As shown in Fig. 1(c), it reveals substantial performance gaps between models, persistent discrepancies with human reasoning, and limitations of existing spatial benchmarks in urban cross-view reasoning. Beyond evaluation, CityCube is further split into training and testing sets. We fine-tune Qwen3-VL of variant scales on the training set using parameter-efficient LoRA, resulting in CityBot-2B, 4B and 8B. Experimental results show the potential of fine-tuning on CityCube in enhancing spatial reasoning, even for relatively small model scales.

In summary, our contributions are threefold:

- **Dataset:** We introduce CityCube, a comprehensive dataset dedicated to cross-view spatial reasoning in urban embodied environments, covering 18.1K images from diverse perspectives across a wide spectrum of urban scenes. And this work further rearranges the imagery through four observation behavior primitives to mimic camera movements.

- **Benchmark:** We build a challenging CvSI benchmark including 5.0K QA pairs across 59 tasks under five fundamental cognition categories. The queries cover three kinds of relation expressions. Based on these problems, this work conducts a comprehensive evaluation over a diverse set of VLMs.

- **Findings:** We uncover several key findings, not limited to: (i) significance of the benchmark, on which leading proprietary and open-source VLMs exhibit lower than 54.1% accuracy, substantiating the challenge of CvSI, (ii) and disparity between current VLMs and human-like reasoning from correlation analyses and case studies.

2 Related Work

VLM Spatial Intelligence Benchmark VLMs have demonstrated significant potential in depth estimation and spatial cognition on spatial intelligence benchmarks like VSI-Bench (Yang et al., 2025a; Cai et al., 2025a; Cheng et al., 2024; Gholami et al., 2025). However, existing benchmarks (Yin et al., 2025) often overlook the inherent cross-view nature of urban environments, limiting their evaluation scope to a broader applications. While prior works such as ViewSpatial (Li et al., 2025), UrbanVideo (Zhao et al., 2025) and Urbench (Zhou et al., 2025) have explored embodied urban spaces to some extent, they primarily focus on single or restricted perspectives, such as bird's-eye views.

Table 1: Comparison of the proposed and popular benchmarks for multi-view spatial intelligence. "/" indicates information not mentioned or included, "+" represents sequential operation. "Total Tasks" represents the amount of the task type. ***Abbreviations*– *Ped.**: Pedestrian, **Veh.**: Vehicle, **Sat.**: Satellite, **Tem.**: Templates, **Ego.**: Egocentric, **Allo.**: Allocentric, **Exo.**: Exocentric.*

Benchmark	Platform	Camera Orientation	QA Num.	Reasoning Annotation	Annotation	Environment	Embodied Questions	Cross Scale	CvSI Categories	Total Tasks
All-Angles (Yeh et al., 2025)	Pedestrian	Ego	2.1K	✗	Human	Indoor	✓	✗	SR, PT	6
MMSI Bench (Yang et al., 2025b)	Pedestrian	Ego, Exo	1K	✓	Human	Indoor	✓	✗	SR, MR, PT	50
ViewSpatial Bench (Li et al., 2025)	Pedestrian	Ego, Exo	5.7K	✗	Rules+Tem.	Indoor & Web images	✓	✗	SR, PT	5
MindCube (Yin et al., 2025)	Pedestrian	Ego, Exo	21.1K	✓	Rules+Tem.	Indoor	✓	✗	MR, PT	5
Ego3D Bench (Gholami et al., 2025)	Vehicle	Ego	8.6K	✓	Rules+Tem.	Outdoor driving	✓	✗	WK	10
OmniSpatial (Jia et al., 2025)	Ped./Vehicle	Ego	1.5K	✓	Human	Web images	✗	✗	PT	50
UrbanFeel (He et al., 2025)	Pedestrian	Ego	14.3K	✗	VFM+Human	Street views	✗	✗	/	11
Urbench (Zhou et al., 2025)	Ped./Satellite	Ego, Allo	11.6K	✗	Rules+LLM+Human	Satellite+Street	✗	✓	/	14
CityCube	**Multi-source** (Ped., Veh., Drone, Sat.)	**Multi-centric** (Ego, Exo, Allo)	**5.0K**	✓	Tem.+LLM+Human	Urban aerial & street views	✓	✓	SR, PT, MR, CR, WK	**59**

As shown in Tab 1, *CityCube provides a unique and comprehensive benchmark by integrating multi-source imagery with multi-centric perspective modeling, and supports the richest set of task types with well-annotated reasoning processes.*

Urban Visual Question Answering Urban Visual Question Answering (VQA) serves as a critical bridge between multi-modal learning and urban informatics. Unlike traditional urban QA systems that rely on retrieving structured information from static databases (Feng et al., 2025a, 2024), urban VQA emphasizes active perception through multi-modal cues, including vision and language. Existing works focus on perception at different scales. As shown in Tab 1, tasks at the macro level encompass geo-spatial querying, urban governance (Zhou et al., 2025; Feng et al., 2025b), and socioeconomic analysis (He et al., 2025; Liu et al., 2025a; Hao et al., 2024). At the micro level, the focus shifts to agentic perception, including object localization (Zhang et al., 2025) and motion planning (Zhao et al., 2025). *However, current benchmarks still exhibit limitations in evaluating multi-image understanding, particularly for tasks requiring advanced cognitive abilities like cross-view reconstruction and perspective transformation in embodied contexts.*

3 CityCube Benchmark

3.1 Overview

As depicted in Fig. 1, CityCube targets multi-scale spatial reasoning under partial observability and dynamical urban viewpoints. This benchmark is built on extensive real-world urban images (collected from 74 cities across the globe, including Singapore and Boston) and virtual images from 2 high-fidelity urban simulators (i.e. EmbodiedCity, and MatrixCity). In total, it offers **18.1K observation points** spanning diverse viewpoints, scales, and scene compositions. Building on this image pool, we carefully curate **5.0K QA pairs** to form the dataset, as shown in Fig. 1(c).

Specifically, the benchmark are structured into four well-designed dimensions, as shown in Fig. 2. The images are rearranged four different observation behavior primitives (**dim1**) with three camera perspectives (**dim2**). On this basis, we further identify five critical task categories (**dim3**) and define three spatial relations (**dim4**) for different queris to assess the VLMs from different CvSI abilities. Structurally, the first two determine the visual perspective and content while collecting the images, while the latter two dimensions are manifested in the textual QA pairs. Together, these variables configure the multi-image QA instances, effectively modeling the evaluation of high-level abilities as a superposition of targeted dimensions. The dataset statistics are illustrated in Appendix B.1.

3.2 Dimension 1: Viewpoint Dynamics

To simulate active perception, we classify collected images to a set of behavior primitives: **(1) Rotation:** hovers and observes at a fixed position, changing only the camera orientation (yaw/pitch). **(2) Orbit:** surrounds a landmark or interested region, including planar and volumetric motions. **(3) Ego-Allocentric View:** simulates cognitive alignment from first-person views (egocentric) to third-person perspectives (allocentric). **(4) Dynaminc Translation:** mimics coarse-to-fine movements, such as zooming in from a skyscraper to a billboard. It involves multiple scales of urban space.

3.3 Dimension 2: View Perspectives

To ensure robustness across diverse embodiments, we standardize three acquisition protocols: **(1) Ground-level Panorama** represents views from ground vehicles (e.g., autonomous cars), covering discrete directions (front, rear, left, right, and diag-

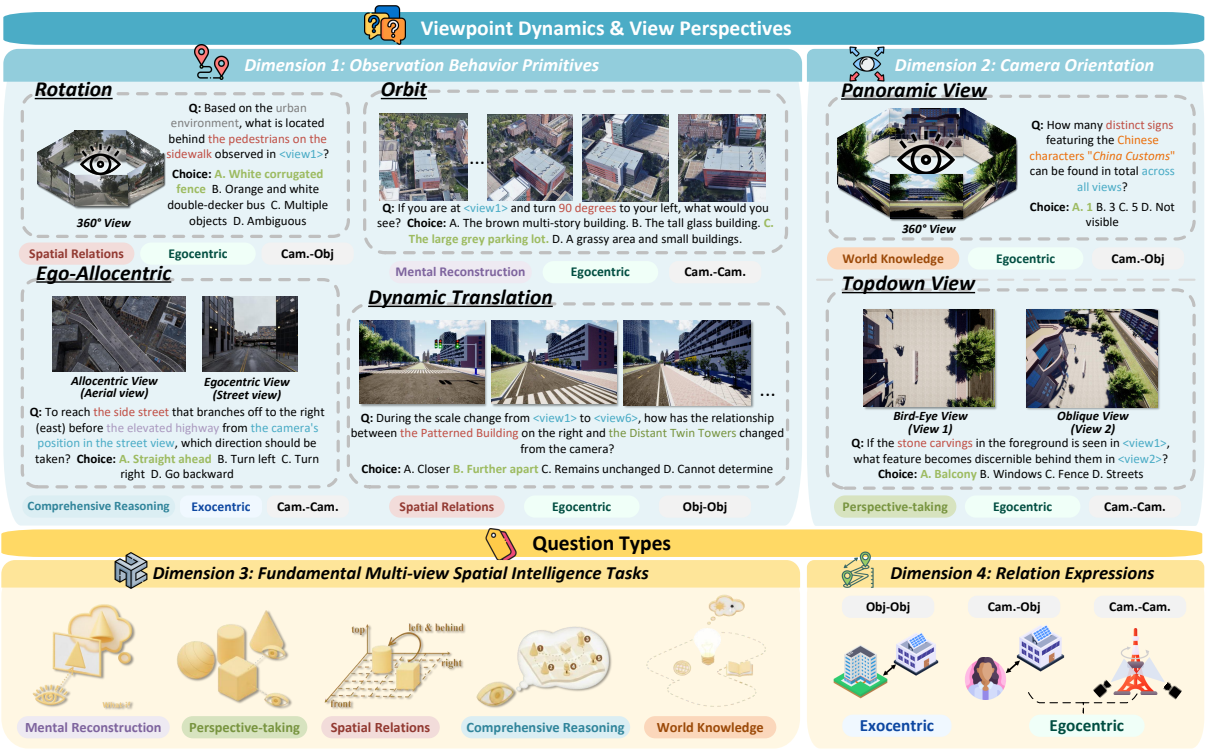


Figure 2: The systematic evaluation protocol of the CityCube benchmark. **Upper Left:** Dim 1 evaluates observation with four representative behavior; **Upper Right:** Dim 2 tests model across various camera orientations; **Bottom Left:** Dim 3 categorizes 59 tasks into 5 fundamental categories; and **Bottom Right:** Dim 4 labels the QA pairs with spatial reference frames.

onals) (Gholami et al., 2025). **(2) Low-altitude Oblique Imagery** represents views from aerial agents like drones, commonly used in 3D urban reconstruction to capture structural facades. **(3) Bird-Eye View** represents a global perspective with minimal occlusion, typical of satellite imagery or high-altitude mapping.

3.4 Dimension 3: Benchmark Tasks

CityCube establishes five fundamental CvSI task categories, built upon existing multi-view reasoning benchmarks in Tab 1 while extending coverage to underexplored dimensions (Cai et al., 2025c):

(1) Mental Reconstruction requires VLMs to infer spatial transformation between views by mentally simulating hypothetical movement through the environment.

(2) Perspective Taking assesses spatial consistency maintenance of VLMs, focusing on cross-view object grounding and inferring relationship shifts as viewpoints change.

(3) Spatial Relation targets the precise quantification of estimating the distance, direction, and topological relations between objects observed across different viewpoints.

(4) Comprehensive Reasoning demands **multi-step spatial inference** for a hypothesis testing. For example, hypothetical navigation requires VLMs to mentally execute actions starting from a specific view and verify the predicted targets referencing other observed views.

(5) World Knowledge probes urban commonsense of VLMs, such as object geometry, affordance, and visibility. Beyond that, we also design challenges for recalling co-visible landmarks, object counting and scene captions across views.

3.5 Dimension 4: Spatial Relation Expressions

To unify spatial reference frames in QA expressions, CityCube tasks utilize three kind of binary spatial relations: **(1) Object-to-Object (Exocentric):** textual geometric relationships between external objects viewed from a spectator perspective. **(2) Camera-to-Object (Egocentric):** spatial relations between the agent's viewpoint and an observed target. **(3) Camera-to-Camera:** relative transformations between two distinct observation positions (e.g., Camera 1 vs. 2).

Rooted in these four critical dimensions, this design facilitates the construction of a systematic,

interpretable, and fine-grained diagnostic benchmark. This benchmark is specifically engineered to model and evaluate high-level spatial reasoning challenges, including multi-view observation integration, geometric consistency under motion, spatial reference frame alignment, and object identity preservation across scale transitions.

4 Benchmark Curation

This section presents how this work constructs the image dataset and benchmark as depicted in Sec. 3, including two main parts: image processing and QA generation.

4.1 Data Collection and Pre-processing

As shown in Fig. 5, we collect urban images that satisfy predefined patterns and explicitly organize them into multi-view sets. The detailed image source are given in the Appendix B.2. In pre-processing, we adopt two distinct pipelines for real-world datasets and 3D simulators, respectively. For real world, we design a multi-step image filtering and matching procedure to construct high-quality multi-view observations. For simulators, we manually collect viewpoint samples, record camera poses and motion trajectories. Across both real-world and simulated data, we ensure that multiple views correspond to the same underlying scene.

Real-world scenes. All real-world images are sourced from public academic datasets. To obtain ground-level panoramic views mentioned in dim 2, we collect images from nuScenes (Caesar et al., 2020), which covers urban environments in Singapore and Boston. Since driving recording contains a large number of duplicate frames, we apply temporal frame skipping and image-similarity deduplication. In addition, GeoText-1652 (Chu et al., 2024) provides satellite and drone imagery captured from arbitrary viewpoints across 72 cities worldwide. For this dataset, we manually design sampling intervals to obtain approximate orbital observations in dim 1 and further remove redundant images based on image similarity. Detailed data sources and sampling strategies are provided in Appendix B.3.

Simulated scenes. To further expand the scale and diversity of the dataset, we collect additional urban imagery from open-source simulators. EmbodiedCity (Gao et al., 2024) is a 3D urban simulator modeled after Beijing, containing multi-scale city elements ranging from large landmarks and commercial buildings to fine-grained objects such as bicycles and billboards. Based on this environment, we manually record agent trajectories over large areas following two representative motion patterns mentioned in dim 1 to support dynamic translation views, such as "taking off" (vertical movement) and "approaching coffee shops" (horizontal movement), which are critical for cross-scale spatial reasoning. Besides, we also supplement the dataset with a set of panoramic views captured by piloting the drone. MatrixCity (Li et al., 2023) is a large-scale aerial-street view dataset from a virtual city. We compute geometric projections using precise camera poses to ensure view consistency between aerial and ground-level images, and manually filter image pairs that satisfy ego-allocentric views in dim 1 and 2. More detailed descriptions are provided in Appendix B.4.

4.2 Question-Answer Generation

To enrich multi-choice QA with CvSI tasks, we feed the processed image sets together with structured context into Gemini-2.5 Pro. The generation relies on three key context engineering strategies:

(1) Contextual Role-Playing: we prompt the VLM to serve as an urban embodiment, along with background knowledge of observation behaviors and perspectives. The specifications of prompts are listed in Sec B.5.1.

(2) Template Coverage: we also provide 59 distinct templates for formatting CvSI tasks, ensuring comprehensive coverage across all structural dimensions of the benchmark.

(3) Geometric Reference Injection: we explicitly supply ground-truth geometric information including camera position, orientation angle and image amount into the context. It is vital for model-based QA generation process, enhancing answer credibility, and mitigating hallucinations that violate physical constraints.

To mitigate potential construction biases arising from model, we implement a rigorous two-stage refinement pipeline as follows:

(1) Blind Filtering: we improve blind filtering (Zhao et al., 2025) for a text-only validation to mitigate textual biases. Specifically, an ensemble of models are used to assess questions without visual input. Questions are scored and stratified into difficulty levels (easy, moderate, hard etc) based on accuracy. By eliminating trivial samples, we obtain a QA dataset with a balanced and hierarchical difficulty distribution.

(2) Human Verification: our annotators verify the rationality and authenticity of the QA pairs, accepting or rejecting entries accordingly. Accepted questions undergo further proofreading to resolve ambiguities, invalid options, or erroneous reasoning, with human-authored reasoning processes added to enrich the annotations. The details are depicted in Sec B.6.2.

5 Experiments

5.1 Evaluation Setups

Evaluated Models. As shown in Tab 2, apart from our fine-tuned models, we also evaluate the performance of 25 models under the CityCube benchmark, including 6 state-of-the-art proprietary models, 16 mainstream open-source models and 3 specialized VLMs trained for spatial reasoning. Notably, the three spatial models are trained on respective spatial intelligence benchmarks, i.e. Spatialvlm (Chen et al., 2024), Omni-spatial (Jia et al., 2025) and SSRL (Liu et al., 2025b).

Evaluation Protocol. Leveraging the multiple-choice format, we compute task-level and overall average accuracy in a straightforward manner. For the human baseline, we recruited two independent groups of ten participants each, all with urban science related academic backgrounds (master's or doctoral students). One group conducted verifications in Sec 4.2, while the other performed the evaluations, ensuring no overlap between the two. More implementation details of evaluation protocol are described in Appendix E.

5.2 Main Results

Our main observations based on the results shown in Table 2 are summarized as follows:

Limited cross-view spatial intelligence across current VLMs. We find that *both* proprietary and open-source VLMs perform poorly on CityCube, indicating that CvSI remains largely unsolved by existing model families. The best-performing proprietary model achieves only 54.1% accuracy, exceeding the strongest open-source model by 9.2%, yet still falling far short of human performance (-34.2%). This consistent gap suggests that the difficulty lies not in model architecture or scale alone, but in the fundamental challenge of multi-view spatial reasoning posed by CityCube.

Effectiveness of fine-tuning on CityCube. Fine-tuning on CityCube consistently improves model performance across all scales. Besides, training with human annotations (e.g., CityBot-4B and -8 B w/ CoT) yields additional gains ($+0.6\%$ to $+3.6\%$), indicating the benefit of reasoning guidance. Notably, the fine-tuned 2B model already surpasses strong proprietary baselines, highlighting the advantage of the benchmark.

Reasoning-oriented VLMs struggle with spatial tasks. Despite their success in math and coding, reasoning-oriented models (e.g., Qwen3-VL-8B-Thinking, Kimi-VL-A3B-Thinking) show no consistent advantage over non-reasoning models on multi-view spatial tasks such as *PT*. This suggests that generic reasoning supervision alone is insufficient to induce strong spatial reasoning in urban environments. We hypothesize that such reasoning requires explicit modeling of view-dependent geometry and spatial transformations, which is not encouraged by current post-training strategies.

CityCube reveals limitations in existing benchmarks. General-purpose spatial models like SpaceOM perform sub-optimally on our urban embodied scenarios. This finding validates the shortcomings of existing benchmarks and emphasizes the unique value of CityCube in evaluating the distinct challenges of CvSI, which are not covered by standard visual question answering evaluations.

Different spatial dimensions exhibit uneven difficulty. Models perform relatively well on *WK* and *CR* tasks. These tasks mainly rely on semantic understanding and logical inference. In contrast, performance drops significantly on *PT*, *SR*, and *MR*. They require precise geometric reasoning and viewpoint transformation across multiple views. The results indicate that current VLMs favor semantic priors over robust spatial representations.

5.3 Task Correlation Analysis

We posit that tasks requiring similar cognitive capabilities will elicit correlated model performance. To analyze the structural relationships among CvSI tasks, we compute the Pearson correlation matrixover 59 tasks evaluated on 25 baseline VLMs. As shown in Fig. 3, we report three key findings:

Strong correlations across task categories. At the dimension level, we observe generally substantial correlations across the five CvSI categories, indicating that spatial intelligence is not naturally decomposed into independent modules. Among all pairs, *MR* and *PT* exhibit the highest inter-dimension correlation ($r = 0.536$), suggesting a shared reliance on underlying cognitive mechanism.

Table 2: Accuracy of 33 VLMs on overall QA pairs. **Only three selected tasks are displayed for each CvSI category besides the overall accuracy due to space limitations.** The best performing model in each category is highlighted **in-bold**, while the second-best is underlined.

Method	Rank	Avg.	World Knowledge				Perspective Taking				Spatial Relation				Mental Recon.				Comp. Reasoning			
			Overall Acc.	*Urban Service*	*Object Ident.*	*Object Counting*	*Overall Acc.*	*Another-view Dir.*	*Third Person View*	*Reverse View*	*Overall Acc.*	*Object Direction*	*Relative Pos.*	*Camera Move.*	*Overall Acc.*	*Multi-Obj View*	*Rotation Pred.*	*Left-turn Pred.*	*Overall Acc.*	*Route Planning*	*Target Direction*	*Location Type*
Baseline																						
Random	-	22.8	19.2	24.0	3.19	24.4	20.5	18.3	21.9	11.5	25.5	25.0	25.2	25.0	22.0	25.1	15.0	24.2	25.2	16.0	28.1	28.6
Human Level	-	88.3	78.6	85.0	84.0	73.2	87.4	93.0	94.2	96.5	90.2	79.2	84.9	86.4	92.4	84.2	89.4	96.8	93.1	100.0	91.5	86.4
Proprietary Models																						
GPT-5.1-251113	3	53.4	**58.3**	47.0	70.2	**53.1**	46.9	32.2	27.7	44.3	**51.6**	56.7	**44.5**	**42.7**	53.7	42.9	**52.2**	38.2	57.8	14.0	**59.3**	48.6
Gemini-2.5-Pro	2	53.8	57.9	**55.0**	57.5	47.0	50.9	33.9	41.3	51.3	50.9	**60.8**	39.5	30.1	52.0	41.4	43.4	**43.4**	59.3	24.0	49.2	55.7
Qwen-3-VL-Plus	5	45.2	40.8	42.0	63.8	37.8	44.6	**47.0**	41.7	**52.2**	46.5	60.0	35.3	31.8	37.1	**44.2**	45.1	37.1	56.7	18.0	47.5	52.1
Step-1o-turbo-vision	4	51.8	55.9	42.0	41.3	46.0	48.2	39.1	35.5	47.8	45.8	56.7	39.5	34.6	52.5	41.9	45.1	37.1	**59.5**	10.0	55.9	**57.1**
Doubao-seed1.6-251015	1	**54.1**	57.7	43.0	**73.4**	48.4	**56.3**	39.1	**51.7**	**52.2**	46.9	55.0	38.7	**35.9**	**54.8**	43.3	47.8	36.6	58.8	**28.0**	55.9	55.7
Skywork-R1V4-Lite	6	40.1	38.6	49.0	43.6	34.3	35.8	27.0	18.2	46.9	34.6	35.8	29.4	19.6	42.9	31.6	43.4	29.6	51.0	8.0	39.0	48.6
Open-source Models																						
Qwen3-VL-8B-Instruct	3	43.1	36.1	20.0	28.7	14.1	37.6	24.4	27.3	25.7	**45.8**	40.0	30.3	37.3	44.2	**37.2**	38.1	34.4	49.6	22.0	44.1	30.0
Qwen3-VL-8B-Thinking	9	39.7	41.8	22.0	41.5	35.7	36.3	28.7	22.7	20.4	39.0	35.8	21.9	36.4	39.0	20.9	36.3	30.7	42.9	18.0	35.6	22.1
GLM-4.1V-9B-Base	5	42.6	39.7	19.0	36.2	24.4	36.2	32.2	28.5	31.0	43.0	**42.5**	33.6	38.6	42.4	28.4	36.3	**37.6**	51.3	32.0	44.1	32.1
GLM-4.1V-9B-Thinking	1	**44.9**	**45.8**	**25.0**	36.2	**40.9**	42.8	**43.5**	24.8	**54.0**	44.8	36.7	28.6	39.1	40.2	33.5	39.8	28.5	51.5	26.0	39.0	27.1
Kimi-VL-A3B-Instruct	10	39.7	36.1	**25.0**	12.8	23.5	33.9	27.8	24.0	45.1	36.4	35.8	24.4	35.0	43.8	28.4	40.7	35.5	49.3	10.0	39.0	23.6
Kimi-VL-A3B-Thinking	13	36.0	32.6	18.0	34.0	27.2	31.6	22.6	19.8	27.4	34.6	25.0	26.1	21.8	39.0	25.1	**43.4**	29.0	42.1	12.0	42.4	27.1
MiMo-VL-7B-SFT	8	40.2	36.9	23.0	30.9	13.2	39.8	32.2	28.9	37.2	38.1	21.7	21.0	37.3	38.9	32.1	31.9	26.3	48.4	16.0	44.1	23.6
MiMo-VL-7B-RL	7	40.9	38.2	21.0	33.0	16.9	39.9	35.7	30.2	30.1	38.4	29.2	26.9	37.3	41.2	34.0	31.9	30.1	48.2	18.0	47.5	23.6
MiniCPM-V-4.5	2	43.9	37.1	20.0	19.2	19.3	**44.3**	40.0	**34.3**	33.6	42.6	36.7	26.9	35.0	43.5	35.8	38.1	31.2	52.5	26.0	37.3	27.1
Ovis2.5-9B	4	42.7	40.7	20.0	30.9	24.9	41.1	36.5	33.1	38.1	39.3	20.8	22.7	45.5	40.9	24.7	39.8	31.2	**53.2**	**40.0**	47.5	28.6
LLaVA-NeXT-Video-7B	16	28.3	32.6	**25.0**	42.6	23.0	21.5	25.2	25.2	8.0	25.9	32.5	18.5	23.2	26.3	23.7	26.6	23.1	36.6	20.0	35.6	29.3
LLaVA-Onevision-7B	6	42.3	37.4	22.0	28.7	27.2	34.8	28.7	25.2	48.7	43.2	30.8	28.6	**50.5**	**45.6**	24.7	37.2	37.1	49.6	20.0	30.5	29.3
InternVL2.5-8B	12	38.7	36.1	16.0	45.7	14.1	31.7	17.4	18.6	44.3	37.2	34.2	33.6	33.2	40.4	24.2	40.7	**37.6**	48.2	10.0	**50.9**	25.0
Skywork-VL-Reward-7B	14	33.2	36.8	5.0	41.5	28.2	34.9	27.0	24.4	36.3	22.8	2.5	21.0	21.8	32.7	21.9	23.9	19.4	44.6	28.0	35.6	19.3
Molmo-7B-D-0924	11	38.7	33.3	17.0	25.5	30.5	33.5	31.3	20.7	30.1	36.8	28.3	**34.5**	32.3	41.8	26.5	33.6	28.0	48.2	26.0	44.1	**33.6**
Phi-4-multimodal-instruct	15	32.0	31.8	23.0	**46.8**	16.4	29.0	33.9	17.8	40.7	27.3	11.7	16.8	25.0	32.0	28.4	15.0	24.7	42.0	18.0	32.2	17.1
Spatial Models																						
Spatial-SSRL-4B	1	**39.8**	**39.2**	22.0	35.1	15.0	31.2	31.3	26.5	21.2	**41.6**	30.0	**31.1**	**47.3**	37.8	**30.2**	35.4	24.7	48.0	**34.0**	**49.2**	27.1
SpaceOm-4B	2	38.9	38.6	**25.0**	**40.4**	**23.5**	**34.3**	**33.9**	23.1	48.7	37.1	**35.0**	19.3	33.2	37.9	26.5	**36.3**	34.9	47.2	8.0	45.8	**35.0**
SpaceThinker-3B	3	38.7	35.8	21.0	38.3	19.3	34.2	25.2	**29.8**	**55.8**	35.6	25.0	29.4	30.9	**40.5**	29.8	33.6	**36.0**	**48.5**	18.0	33.9	33.6
Fine-Tuning: Test set																						
Qwen3-VL-2B (before)	9	30.4	26.4	30.0	10.0	22.7	23.4	25.0	12.0	33.3	37.1	50.0	41.7	31.8	25.4	13.6	16.7	26.3	36.7	20.0	50.0	28.6
CityBot-2B (CoT)	4	60.2	61.5	40.0	**70.0**	59.1	**62.8**	48.3	**56.0**	83.3	60.1	**75.0**	41.7	45.5	50.9	31.8	38.3	31.6	67.4	60.0	50.0	57.1
CityBot-2B (w/o CoT)	6	55.8	58.2	40.0	50.0	68.2	50.0	41.7	28.0	75.0	57.3	66.7	41.7	45.5	46.1	18.2	33.3	42.1	67.4	40.0	50.0	60.0
Qwen3-VL-4B (before)	8	36.6	38.5	20.0	60.0	40.9	27.7	41.7	28.0	12.5	37.8	50.0	25.0	36.4	39.2	27.3	41.7	15.8	38.8	40.0	33.3	28.6
CityBot-4B (CoT)	2	61.0	**67.0**	**50.0**	**70.0**	72.7	59.6	**50.0**	48.0	66.7	**58.1**	66.7	41.7	50.0	54.9	27.3	41.7	**68.4**	67.4	**80.0**	50.0	**71.4**
CityBot-4B (w/o CoT)	3	60.4	62.6	34.3	50.0	68.2	53.2	33.3	28.0	75.0	58.0	66.7	41.7	45.5	**59.8**	31.8	41.7	52.6	**69.4**	40.0	**83.3**	64.3
Qwen3-VL-8B (before)	7	37.1	40.7	20.0	60.0	45.5	26.6	8.3	16.0	58.3	42.0	66.7	33.3	31.8	38.2	22.7	41.7	26.3	35.7	40.0	50.0	57.1
CityBot-8B (CoT)	1	**61.4**	64.8	**50.0**	**70.0**	**77.3**	**62.8**	41.7	52.0	**91.7**	58.0	**75.0**	**50.0**	**54.6**	52.9	**40.9**	**50.0**	36.8	64.8	60.0	66.7	**71.4**
CityBot-8B (w/o CoT)	5	57.8	58.2	40.0	50.0	68.2	54.3	**50.0**	32.0	83.3	57.3	**75.0**	33.3	45.5	53.9	31.8	25.0	52.6	65.3	20.0	66.7	50.0

Dense correlations across different tasks. At a finer granularity, task-level analysis reveals dense cross-category correlations, with 81.4% of high-correlation pairs spanning different CvSI dimensions. For example, *Behind View* (T33, SR) and *Left-turn Prediction* (T15, CR) exhibit near-perfect correlation ($r = 0.954$), indicating that panoramic spatial perception is tightly coupled with dynamic viewpoint reasoning. Similarly, strong associations are observed between cross-scale semantic tasks such as *Hierarchical Structure* (T3, WK) and *Object Localization Sequence* (T8, CR) ($r = 0.965$), reflecting synchronized semantic–geometric reasoning under dynamic scale changes.

Metric estimation as a weakly correlated capability. Metric estimation shows weaker coupling with other tasks. For example, *Height Estimation* (T40) shows negligible correlation with most other tasks, suggesting that precise metric reasoning constitutes a distinct capability, weakly linked to view-dependent or semantic spatial reasoning processes.

5.4 The VLMs Error Analysis

Through a case-by-case qualitative investigation (as depicted in Appendix D), we identify four primary failure modes of current VLMs in urban spatial reasoning:

Limited sensitivity to small-scale urban entities. In complex urban environments, models frequently fail to recognize small-scale urban en-

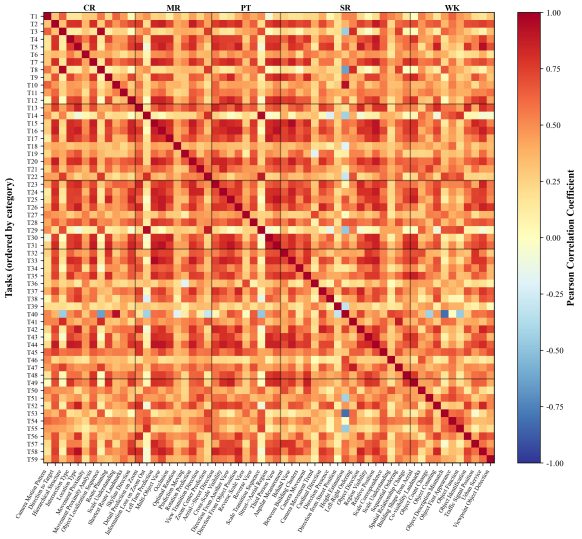


Figure 3: Task correlation matrix. Each axis corresponds to the 59 tasks after classification, and color intensity indicates the strength of correlations.

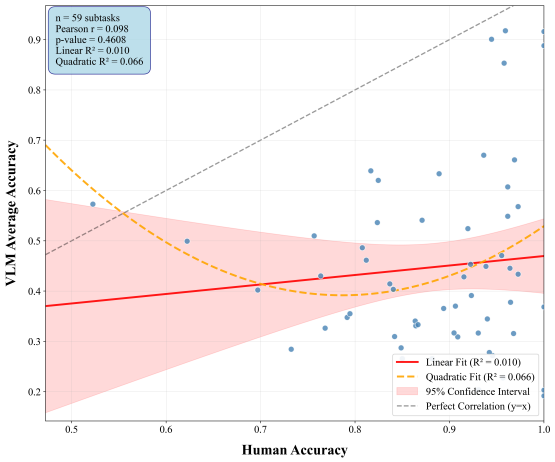


Figure 4: Human vs AI task correlation. The scatters illustrate the performance correlation across 59 tasks, where each individual point represents a specific task.

tities or entirely overlook critical landmarks. This issue is particularly pronounced in scenes with high semantic density, where numerous visually similar objects compete for attention.

Failures in egocentric spatial reasoning. Models struggle to correctly predict changes in relative position, orientation, and visibility when the egocentric viewpoint shifts. This failure in "mental rotation" indicates a lack of a reliable internal spatial representation.

Insufficient cross-view consistency. Models exhibit difficulty in establishing correct correspondences across perspectives. For instance, models often mismatch a building in street imagery with an unrelated structure in aerial views, leading to erroneous reasoning about spatial relationships and world knowledge.

Misinterpretation of motion and scale dynamics. Models struggle to interpret the camera movements (e.g., forward/backward translation or directional turns) and the resulting dynamic scale changes. Such failures in motion understanding directly impede the ability to perform complex scene reconstruction and navigation-related reasoning.

5.5 Human-AI Difficulty Bias Analysis

To examine whether the difficulty of spatial tasks for humans aligns with VLM performance, we analyze the correlation between human baseline accuracy and the average performance of evaluated VLMs across 59 subtasks. Fig. 4 reveals a Pearson correlation coefficient of $r = 0.098$ with a p-value of 0.4608. This extremely low correlation ($R^2 = 0.010$) indicates that tasks found difficult by VLMs do not necessarily pose a challenge for human, and vice versa. Notably, the AI performance distribution is highly dispersed (range: $[19.2\%, 91.8\%]$), suggesting that VLMs might merely sensitive to low-level visual features rather than spatial mental modeling. This divergence confirms that CityCube captures unique spatial challenges that are non-trivial for current model architectures, despite being intuitive for humans.

6 Conclusion

In this paper, we presented **CityCube**, a comprehensive benchmark specifically designed to evaluate CvSI of VLMs in urban environments. CityCube encompasses 59 tasks across 5 cognitive categories, supported by a large-scale collection of images from dynamic viewpoints and diverse orientations. The resulting dataset contains 5,022 multiple-choice questions (MCQs), each rigorously annotated and verified by humans.

Our extensive evaluation of 33 VLMs reveals that CvSI remains extremely challenging, even for very large-scale models, "thinking" models and specialized spatial models. While our fine-tuned **CityBot** models (based on 2B, 4B and 8B backbones) outperform leading proprietary models, a substantial gap to human-level spatial cognition persists. We hope that CityCube can serve as a foundation for future studies on spatially grounded learning paradigms and as a diagnostic tool for developing next-generation VLMs with stronger urban spatial intelligence.

7 Limitations

This work has several limitations. First, although CityCube covers diverse cross-view spatial tasks, we do not explicitly isolate perspective-induced cognitive biases, such as viewpoint asymmetry and reference-frame ambiguity, which may systematically influence spatial reasoning performance. Second, CityCube is constructed in a simulated urban environment; the impact of Sim-to-Real transfer and domain gaps is not evaluated in this study. Third, our analysis focuses on task-level performance and does not probe internal representations or multi-view fusion mechanisms, limiting interpretability of model failure modes. Finally, while we hypothesize the importance of explicit spatial supervision, we do not implement or validate dedicated post-training strategies (e.g., spatial Chain-of-Thought) in this work. Future work will address these issues by analyzing viewpoint biases, extending evaluation to Sim2real settings, and exploring spatially grounded architectural and post-training designs.

8 Ethics Statement

This research exclusively utilizes publicly available datasets, programs, and pre-trained models. All data annotation was conducted with informed consent from participants, and the datasets do not contain information that could compromise individual privacy or public safety. All procedures strictly adhere to the guidelines established by the ACL Code of Ethics. Therefore, this work does not raise any ethical concerns.

References

Panos Achlioptas, Ahmed Abdelreheem, Fei Xia, Mohamed Elhoseiny, and Leonidas Guibas. 2020. Referit3d: Neural listeners for fine-grained 3d object identification in real-world scenes. In *European conference on computer vision*, pages 422–440. Springer.

Daichi Azuma, Taiki Miyanishi, Shuhei Kurita, and Motoaki Kawanabe. 2022. Scanqa: 3d question answering for spatial scene understanding. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19129–19139.

Shuai Bai, Yuxuan Cai, and Ke Zhu. 2025. Qwen3-vl technical report. *Preprint*, arXiv:2511.21631.

Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. 2020. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11621–11631.

Wenxiao Cai, Iaroslav Ponomarenko, Jianhao Yuan, Xiaoqi Li, Wankou Yang, Hao Dong, and Bo Zhao. 2025a. Spatialbot: Precise spatial understanding with vision language models. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pages 9490–9498. IEEE.

Zhongang Cai, Ruisi Wang, and Lei Yang. 2025b. Scaling spatial intelligence with multimodal foundation models. *arXiv preprint arXiv:2511.13719*.

Zhongang Cai, Yubo Wang, Qingping Sun, Ruisi Wang, Chenyang Gu, Wanqi Yin, Zhiqian Lin, Zhitao Yang, Chen Wei, Xuanke Shi, and 1 others. 2025c. Has gpt-5 achieved spatial intelligence? an empirical study. *arXiv preprint arXiv:2508.13142*, 3.

Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niessner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. 2017. Matterport3d: Learning from rgb-d data in indoor environments. *arXiv preprint arXiv:1709.06158*.

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

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

Meng Chu, Zhedong Zheng, Wei Ji, Tingyu Wang, and Tat-Seng Chua. 2024. Towards natural language-guided drones: Geotext-1652 benchmark with spatial relation matching. In *European Conference on Computer Vision*, pages 213–231. Springer.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. 2017. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5828–5839.

Jingtao Ding, Yunke Zhang, Yu Shang, Yuheng Zhang, Zefang Zong, Jie Feng, Yuan Yuan, Hongyuan Su, Nian Li, Nicholas Sukiennik, and 1 others. 2025.

Understanding world or predicting future? a comprehensive survey of world models. *ACM Computing Surveys*, 58(3):1–38.

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

SM Ali Eslami, Danilo Jimenez Rezende, Frederic Besse, Fabio Viola, Ari S Morcos, Marta Garnelo, Avraham Ruderman, Andrei A Rusu, Ivo Danihelka, Karol Gregor, and 1 others. 2018. Neural scene representation and rendering. *Science*, 360(6394):1204–1210.

Jie Feng, Tianhui Liu, Yuwei Du, Siqi Guo, Yuming Lin, and Yong Li. 2025a. Citygpt: Empowering urban spatial cognition of large language models. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pages 591–602.

Jie Feng, Shengyuan Wang, Tianhui Liu, Yanxin Xi, and Yong Li. 2025b. Urbanllava: A multi-modal large language model for urban intelligence with spatial reasoning and understanding. *arXiv preprint arXiv:2506.23219*.

Jie Feng, Jun Zhang, Junbo Yan, Xin Zhang, Tianjian Ouyang, Tianhui Liu, Yuwei Du, Siqi Guo, and Yong Li. 2024. Citybench: Evaluating the capabilities of large language model as world model. *arXiv e-prints*, pages arXiv–2406.

Rao Fu, Jingyu Liu, Xilun Chen, Yixin Nie, and Wenhan Xiong. 2024. Scene-llm: Extending language model for 3d visual understanding and reasoning. *arXiv preprint arXiv:2403.11401*.

Chen Gao, Baining Zhao, Weichen Zhang, Jinzhu Mao, Jun Zhang, Zhiheng Zheng, Fanhang Man, Jianjie Fang, Zile Zhou, Jinqiang Cui, and 1 others. 2024. Embodiedcity: A benchmark platform for embodied agent in real-world city environment. *arXiv preprint arXiv:2410.09604*.

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

Dong Guo, Faming Wu, and Zuquan Song. 2025. Seed1.5-vl technical report. *Preprint*, arXiv:2505.07062.

Zoey Guo, Yiwen Tang, Ray Zhang, Dong Wang, Zhigang Wang, Bin Zhao, and Xuelong Li. 2023. Viewrefer: Grasp the multi-view knowledge for 3d visual grounding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15372–15383.

Xixuan Hao, Wei Chen, Yibo Yan, Siru Zhong, Kun Wang, Qingsong Wen, and Yuxuan Liang. 2024. Urbanvlp: A multi-granularity vision-language pretrained foundation model for urban indicator prediction. *CoRR*.

Jun He, Yi Lin, Zilong Huang, Jiacong Yin, Junyan Ye, Yuchuan Zhou, Weijia Li, and Xiang Zhang. 2025. Urbanfeel: A comprehensive benchmark for temporal and perceptual understanding of city scenes through human perspective. *arXiv preprint arXiv:2509.22228*.

Yining Hong, Haoyu Zhen, Peihao Chen, Shuhong Zheng, Yilun Du, Zhenfang Chen, and Chuang Gan. 2023. 3d-llm: Injecting the 3d world into large language models. *Advances in Neural Information Processing Systems*, 36:20482–20494.

Jiangyong Huang, Silong Yong, Xiaojian Ma, Xiongkun Linghu, Puhao Li, Yan Wang, Qing Li, Song-Chun Zhu, Baoxiong Jia, and Siyuan Huang. 2024. An embodied generalist agent in 3d world. In *Proceedings of the 41st International Conference on Machine Learning*, pages 20413–20451.

Mengdi Jia, Zekun Qi, Shaochen Zhang, Wenyao Zhang, Xinqiang Yu, Jiawei He, He Wang, and Li Yi. 2025. Omnispatial: Towards comprehensive spatial reasoning benchmark for vision language models. *arXiv preprint arXiv:2506.03135*.

Dingming Li, Hongxing Li, Zixuan Wang, Yuchen Yan, Hang Zhang, Siqi Chen, Guiyang Hou, Shengpei Jiang, Wenqi Zhang, Yongliang Shen, and 1 others. 2025. Viewspatial-bench: Evaluating multi-perspective spatial localization in vision-language models. *arXiv preprint arXiv:2505.21500*.

Yixuan Li, Lihan Jiang, Linning Xu, Yuanbo Xiangli, Zhenzhi Wang, Dahua Lin, and Bo Dai. 2023. Matrixcity: A large-scale city dataset for city-scale neural rendering and beyond. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3205–3215.

Tianhui Liu, Jie Feng, Hetian Pang, Xin Zhang, Tianjian Ouyang, Zhiyuan Zhang, and Yong Li. 2025a. Citylens: Benchmarking large language-vision models for urban socioeconomic sensing. *arXiv preprint arXiv:2506.00530*.

Yuhong Liu, Beichen Zhang, Yuhang Zang, Yuhang Cao, Long Xing, Xiaoyi Dong, Haodong Duan, Dahua Lin, and Jiaqi Wang. 2025b. Spatial-ssrl: Enhancing spatial understanding via self-supervised reinforcement learning. *arXiv preprint arXiv:2510.27606*.

Arjun Majumdar, Anurag Ajay, Xiaohan Zhang, Pranav Putta, Sriram Yenamandra, Mikael Henaff, Sneha Silwal, Paul Mcvay, Oleksandr Maksymets, Sergio Arnaud, and 1 others. 2024. Openeqa: Embodied question answering in the era of foundation models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16488–16498.

OpenAI. 2025. Gpt-5 system card. Accessed: 2026-01-03.

Jean Piaget. 2013. *Child's conception of space: Selected works vol 4*. Routledge.

Zekun Qi, Runpei Dong, Shaochen Zhang, Haoran Geng, Chunrui Han, Zheng Ge, Li Yi, and Kaisheng Ma. 2024. Shapellm: Universal 3d object understanding for embodied interaction. In *European Conference on Computer Vision*, pages 214–238. Springer.

Chan Hee Song, Valts Blukis, Jonathan Tremblay, Stephen Tyree, Yu Su, and Stan Birchfield. 2025. Robospatial: Teaching spatial understanding to 2d and 3d vision-language models for robotics. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 15768–15780.

Barbara Tversky. 2019. *Mind in motion: How action shapes thought*. Basic Books.

Tianwen Wei, Liang Zhao, and Yahui Zhou. 2023. Skywork: A more open bilingual foundation model. *Preprint*, arXiv:2310.19341.

Jihan Yang, Shusheng Yang, Anjali W Gupta, Rilyn Han, Li Fei-Fei, and Saining Xie. 2025a. Thinking in space: How multimodal large language models see, remember, and recall spaces. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 10632–10643.

Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. 2024. Depth anything: Unleashing the power of large-scale unlabeled data. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10371–10381.

Sihan Yang, Runsen Xu, Yiman Xie, Sizhe Yang, Mo Li, Jingli Lin, Chenming Zhu, Xiaochen Chen, Haodong Duan, Xiangyu Yue, and 1 others. 2025b. Mmsi-bench: A benchmark for multi-image spatial intelligence. *arXiv preprint arXiv:2505.23764*.

Chun-Hsiao Yeh, Chenyu Wang, Shengbang Tong, Ta-Ying Cheng, Ruoyu Wang, Tianzhe Chu, Yuexiang Zhai, Yubei Chen, Shenghua Gao, and Yi Ma. 2025. Seeing from another perspective: Evaluating multi-view understanding in mllms. *arXiv preprint arXiv:2504.15280*.

Baiqiao Yin, Qineng Wang, Pingyue Zhang, Jianshu Zhang, Kangrui Wang, Zihan Wang, Jieyu Zhang, Keshigeyan Chandrasegaran, Han Liu, Ranjay Krishna, and 1 others. 2025. Spatial mental modeling from limited views. In *Structural Priors for Vision Workshop at ICCV'25*.

Sha Zhang, Di Huang, Jiajun Deng, Shixiang Tang, Wanli Ouyang, Tong He, and Yanyong Zhang. 2024. Agent3d-zero: An agent for zero-shot 3d understanding. In *European Conference on Computer Vision*, pages 186–202. Springer.

Weichen Zhang, Zile Zhou, Zhiheng Zheng, Chen Gao, Jinqiang Cui, Yong Li, Xinlei Chen, and Xiao-Ping Zhang. 2025. Open3dvqa: A benchmark for comprehensive spatial reasoning with multimodal large language model in open space. *arXiv preprint arXiv:2503.11094*.

Baining Zhao, Jianjie Fang, Zichao Dai, Ziyou Wang, Jirong Zha, Weichen Zhang, Chen Gao, Yue Wang, Jinqiang Cui, Xinlei Chen, and 1 others. 2025. Urbanvideo-bench: Benchmarking vision-language models on embodied intelligence with video data in urban spaces. *arXiv preprint arXiv:2503.06157*.

Baichuan Zhou, Haote Yang, Dairong Chen, Junyan Ye, Tianyi Bai, Jinhua Yu, Songyang Zhang, Dahua Lin, Conghui He, and Weijia Li. 2025. Urbench: A comprehensive benchmark for evaluating large multimodal models in multi-view urban scenarios. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 10707–10715.

Chenming Zhu, Tai Wang, Wenwei Zhang, Jiangmiao Pang, and Xihui Liu. 2024. Llava-3d: A simple yet effective pathway to empowering lmms with 3d-awareness. *arXiv preprint arXiv:2409.18125*.

A Appendix

In the supplementary materials, we provide the following:

- Details of data collection, processing, automatic generation and refinement of **CityCube** Benchmark (Sec B).
- Details of reproduction, including code for QA generation, model training scripts, image datasets, and the detail model information (Sec C).
- Details of VLMs error and their corresponding reason (Sec D).
- Details of experiment setup, including training hyperparameter, dataset settings and human evaluation UI interface (Sec E).
- Further discussion on relative research, especially in 3D QA (Sec F).

B Dataset Generation Pipeline

B.1 Dataset Visualization

We analyzed the distribution of spatial reference frames and spatial task frequencies associated with each observation behavior, with the results summarized in the stacked bar chart in Fig. 6. The training and test sets contain approximately 4.5k and 0.5k questions, respectively. We construct each split using a stratified sampling strategy, ensuring that the proportion of each task type remains consistent with the overall QA distribution of the dataset. We further present word cloud visualizations of the textual content in the training set, as illustrated in Fig. 7.

B.2 Details of Image Acquisition

Our sources span both real-world environments and photorealistic virtual scenes. We leverage real-world sensor datasets—GeoText-1652 (Chu et al., 2024) and nuScenes (Caesar et al., 2020), to acquire raw data for perspectives involving camera behavior primitives of Rotation and Orbit. We further augment the benchmark with extensive first-person to third-person camera motion imagery derived from embodied vision, which are based on virtual environments—EmbodiedCity (Gao et al., 2024) and MatrixCity (Li et al., 2023).

The core datasets are selected for their complementary strengths in providing diverse, high-quality visual data while adhering to rigorous data ethics standards. Key details are outlined below:

- **nuScenes (Caesar et al., 2020)**: This multimodal autonomous driving dataset provides 1,000 large-scale scenes from real-world urban environments (Boston and Singapore). Its synchronized data from six cameras, LiDAR, and radar, along with comprehensive 3D annotations and precise calibration, offers a rich, real-world foundation for studying multi-view geometry and complex camera motions.
- **GeoText-1652 (Chu et al., 2024)**: This benchmark extends real-world imagery with spatial language annotations. Based on the established University-1652 image set, it provides high-resolution ground-level and aerial images. Its key contribution to our work is the precise 360-degree image correspondence the aerial scan of urban buildings, facilitating tasks that require spatial and linguistic grounding.
- **MatrixCity (Li et al., 2023)**: A large-scale photorealistic synthetic dataset built with Unreal Engine 5. It provides over 500,000 street-view and aerial images with pixel-perfect ground-truth information (e.g., camera poses, depth, normal maps) and full control over environmental conditions (weather, lighting). It serves as a primary, privacy-safe source for generating diverse first-person camera motion trajectories.
- **EmbodiedCity (Gao et al., 2024)**: This benchmark platform supports embodied AI agents in city-scale environments. It enables the generation of extensive, realistic first-person visual experience data for navigation and interaction tasks, perfectly aligning with our need for embodied, egocentric visual data in complex virtual urban settings.

Privacy and Ethical Compliance Statement. We strictly adhere to data privacy and ethical research standards. All real-world data utilized in this study (nuScenes and GeoText-1652) are sourced from established, publicly released academic datasets that have undergone formal anonymization and curation processes, involving no collection of personal identifiers. Crucially, a significant portion of our

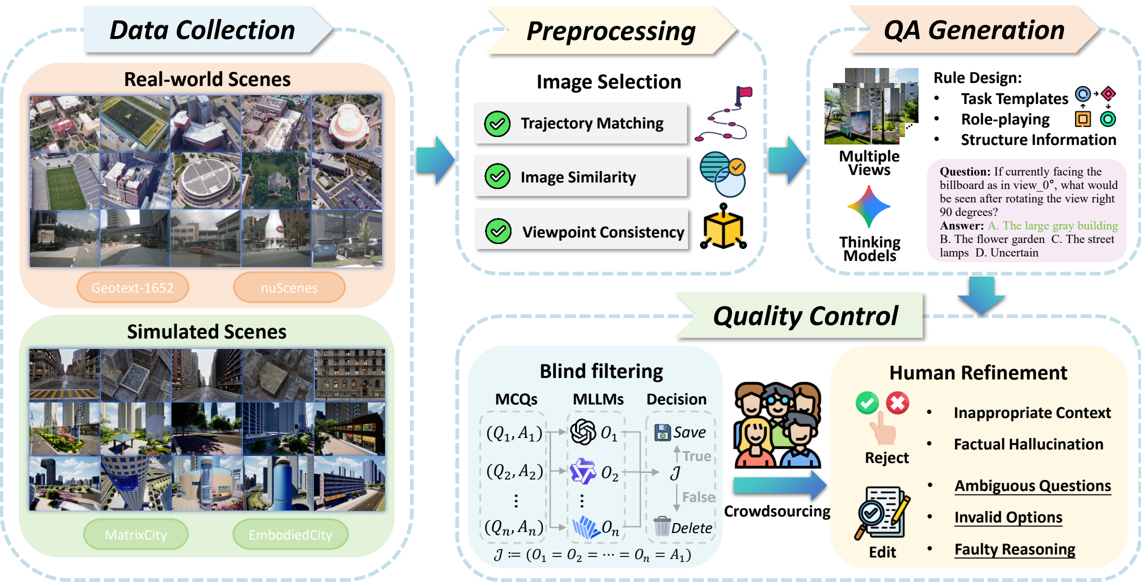


Figure 5: Illustration of the CityCube-Bench construction pipeline. Images are collected from diverse real-world datasets and urban simulators; relevant image sets are carefully selected; complex QA tasks and detailed reasoning processes are annotated through human–AI collaboration; and all data undergo rigorous quality control.

Figure 6: Histogram of different behaviors and relation expressions

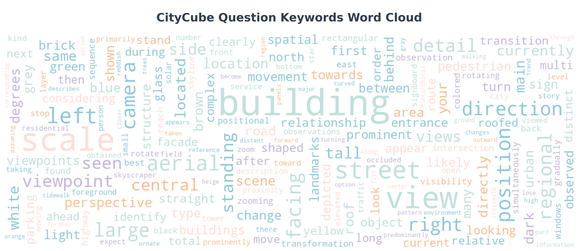


Figure 7: Word cloud of CityCube QA dataset.

training data—particularly for complex, embodied first-person perspectives—is generated from fully synthetic environments (MatrixCity and EmbodiedCity). This strategy inherently eliminates any risk associated with the privacy of individuals, as no real human subjects or private spaces are involved. All data is used strictly for non-commercial academic research in full compliance with their respective licenses.

B.3 Details of Multi-view Images Processing

B.3.1 nuScenes

Due to the high temporal redundancy in continuous driving sequences, directly using all frames would introduce substantial duplication and bias the distribution toward near-identical views. To address this issue, we adopt a two-stage frame reduction strategy: *temporal subsampling* followed by *appearance-based deduplication*. Specifically, frames are first sampled at a fixed temporal interval to reduce redundancy at the sequence level. Then, a lightweight visual similarity test is applied to further filter near-duplicate frames.

For each candidate timestep, synchronized images from six surrounding cameras are aggregated into a compact multi-view representation. The average pixel-wise absolute difference between consecutive aggregated views is used as a proxy for visual change. Frames with appearance variation below a predefined threshold are discarded. This procedure

Algorithm 1: Multi-view Frame Temporal Sampling and Deduplication for nuScenes

Input: Driving dataset $\mathcal{D}$ with scenes and synchronized multi-camera frames; Temporal step size s; similarity threshold τ; optional per-scene limit M

Output: A set of exported multi-view frame groups with metadata

foreach *scene* $\mathcal{S}$ *in* $\mathcal{D}$ **do**
- Retrieve ordered frame list $\{\mathbf{F}_1, \mathbf{F}_2, \ldots, \mathbf{F}_N\}$;
- Initialize $\mathbf{A}_{\text{prev}} \leftarrow \varnothing$, counter $k \leftarrow 0$;
- **for** $i = 1, 1+s, \ldots, N$ **do**
  - **if** M *is set* ***and*** $k \geq M$ **then** **break**
  - Let $\mathbf{F}_i = \{I_i^{(1)}, \ldots, I_i^{(6)}\}$ be six synchronized camera images;
  - **if** *any image in* $\mathbf{F}_i$ *is missing* **then** **continue**
  - Construct concatenate view $\mathbf{A}_i \leftarrow \textsc{Concate}(\mathbf{F}_i)$;
  - **if** $\mathbf{A}_{prev} \neq \varnothing$ **then**
    - Compute appearance difference $d \leftarrow \frac{1}{|\mathbf{A}_i|} \sum |\mathbf{A}_i - \mathbf{A}_{\text{prev}}|$
    - **if** $d < \tau$ **then** **continue**
  - Export all images in $\mathbf{F}_i$ as one multi-view sample;
  - Save associated metadata (scene ID, frame index, camera list);
  - $\mathbf{A}_{\text{prev}} \leftarrow \mathbf{A}_i$;
  - $k \leftarrow k + 1$;

preserves scene diversity while significantly reducing redundant observations. The retained frames are stored as structured multi-view samples, each associated with metadata describing scene identity and camera configuration.

B.3.2 Geotext-1652

Unlike driving datasets with strong temporal continuity, images in GeoText-1652 are organized as viewpoint sequences centered around prominent landmarks, exhibiting systematic variations in camera distance and altitude.

To obtain representative landmark-centric multi-view observations while avoiding redundant samples, we adopt a rule-based sampling strategy with

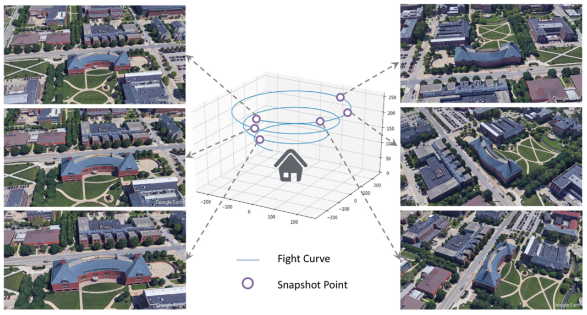


Figure 8: A data example of viewpoint dynamic in Geotext-1652.

manually designed intervals. Specifically, for each landmark sequence, we extract two complementary subsets: (i) *orbit views*, which capture appearance changes under small viewpoint variations at similar altitudes, and (ii) *spiral views*, which emphasize significant altitude and scale changes. These two subsets jointly approximate local and global perceptual variations around a landmark.

Given an ordered image sequence, orbit views are sampled with a short interval to retain fine-grained viewpoint diversity, while spiral views are sampled with a larger interval to highlight elevation changes. To further control redundancy and balance the dataset, we impose an upper bound on the number of retained images per subset. The selected images are exported into separate directories according to their view type, preserving the original filenames for traceability. This strategy yields compact yet diverse multi-view observations suitable for landmark-centric perception and reasoning tasks.

B.3.3 Cross-view Geometric Pairing in MatrixCity

MatrixCity is a large-scale virtual city dataset providing synchronized aerial and street-level imagery with precise camera poses. Unlike real-world datasets where camera geometry may be noisy or partially missing, MatrixCity offers accurate extrinsic parameters, enabling explicit geometric reasoning between ego-centric (street) and allocentric (aerial) views.

To construct reliable aerial-street image pairs, we perform geometry-aware cross-view matching based on camera poses and viewing configurations (as depicted in 3). Given a street-level image, candidate aerial views are filtered and scored through a sequence of geometric constraints. Only image pairs that satisfy both spatial overlap and view-

Algorithm 2: Landmark-centric Multi-view Sampling for GeoText-1652

Input: Image folders $\{\mathcal{F}_1, \ldots, \mathcal{F}_K\}$; orbit sampling step s_o; spiral sampling step s_s; maximum samples per type M
Output: Two subsets per folder: orbit views and spiral views

foreach *folder* $\mathcal{F}$ **do**
- Load ordered image sequence $\{I_1, I_2, \ldots, I_N\}$;
- Initialize orbit index set $\mathcal{I}_o \leftarrow \varnothing$;
- Initialize spiral index set $\mathcal{I}_s \leftarrow \varnothing$;
- **for** $i = 1, 1 + s_o, \ldots, N$ **do**
  - Append i to $\mathcal{I}_o$;
- **for** $i = 1, 1 + s_s, \ldots, N$ **do**
  - Append i to $\mathcal{I}_s$;
- **if** $N \notin \mathcal{I}_o$ **then**
  - Append N to $\mathcal{I}_o$;
- **if** $N \notin \mathcal{I}_s$ **then**
  - Append N to $\mathcal{I}_s$;
- Truncate $\mathcal{I}_o$ and $\mathcal{I}_s$ to at most M elements;
- **foreach** $i \in \mathcal{I}_o$ **do**
  - Export image I_i to orbit subset;
- **foreach** $i \in \mathcal{I}_s$ **do**
  - Export image I_i to spiral subset;

point consistency are retained, while ambiguous or degenerate cases are manually filtered out. This process ensures strong correspondence between ego-centric observations and their allocentric counterparts.

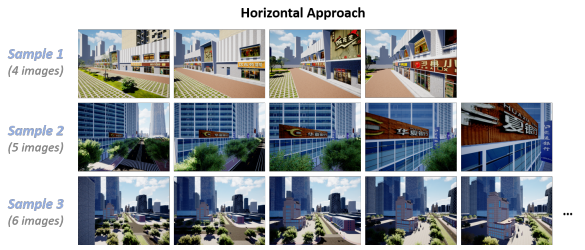
Specifically, the pairing procedure relies on five core geometric factors: (1) horizontal Euclidean distance between camera positions in the ground plane; (2) height difference between aerial and street cameras; (3) overlap of projected viewing areas on the ground; (4) consistency of viewing orientation in the horizontal plane, and (5) ground-plane projection of camera viewing rays. These factors are applied sequentially to prune invalid candidates and to compute a final matching score for pair selection.

We emphasize that the image data does not assume perfect viewpoint control in real-world data. Instead, it focuses on constructing view sets with consistent scene identity and interpretable view-

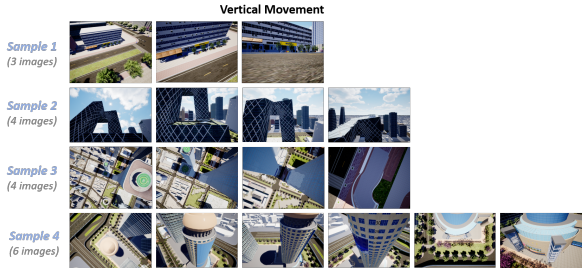
Algorithm 3: Geometry-aware Aerial–Street Pairing for MatrixCity

Input: Street images $\mathcal{S} = \{S_i\}$ with poses; Aerial images $\mathcal{A} = \{A_j\}$ with poses; Height range $[h_{\min}, h_{\max}]$; distance threshold $d_{\max}$; Orientation threshold $\theta_{\max}$
Output: Paired aerial–street image set $\mathcal{P}$

Initialize $\mathcal{P} \leftarrow \varnothing$;
foreach *street image* $S_i \in \mathcal{S}$ **do**
- Extract street camera position $\mathbf{p}_s$ and orientation $\mathbf{o}_s$;
- **if** S_i *violates diversity constraints* **then**
  - continue;
- Compute ground projection $(\mathbf{g}_s, r_s)$ from $(\mathbf{p}_s, \mathbf{o}_s)$;
- Initialize best match $A^\star \leftarrow \varnothing$, best score $c^\star \leftarrow \infty$;
- **foreach** *aerial image* $A_j \in \mathcal{A}$ **do**
  - Extract aerial camera position $\mathbf{p}_a$ and orientation $\mathbf{o}_a$;
  - `// Height filtering`
  - **if** $(\mathbf{p}_a^z - \mathbf{p}_s^z) \notin [h_{\min}, h_{\max}]$ **then**
    - continue;
  - `// Horizontal distance filtering`
  - Compute $d_h \leftarrow \|\mathbf{p}_a^{xy} - \mathbf{p}_s^{xy}\|_2$;
  - **if** $d_h > d_{\max}$ **then**
    - continue;
  - `// Ground projection and viewing overlap`
  - Compute ground projection $(\mathbf{g}_a, r_a)$ from $(\mathbf{p}_a, \mathbf{o}_a)$;
  - Compute viewing center distance $d_g \leftarrow \|\mathbf{g}_a^{xy} - \mathbf{g}_s^{xy}\|_2$;
  - **if** $d_g > r_a + r_s$ **then**
    - continue;
  - `// Cross-view orientation consistency`
  - Compute horizontal orientation angle $\theta \leftarrow \angle(\mathbf{o}_a^{xy}, \mathbf{o}_s^{xy})$;
  - **if** $\theta > \theta_{\max}$ **then**
    - continue;
  - `// Matching score`
  - Compute score $c \leftarrow \alpha d_h + \beta d_g + \gamma(\mathbf{p}_a^z - \mathbf{p}_s^z)$;
  - **if** $c < c^\star$ **then**
    - $c^\star \leftarrow c$, $A^\star \leftarrow A_j$;
- **if** $A^\star \neq \varnothing$ **then**
  - Add pair $(S_i, A^\star)$ to $\mathcal{P}$;
  - Update diversity state;



(a) Multi-scale views in horizontal approach during dynamic translation.



(b) Multi-scale views in vertical movement during dynamic translation.

Figure 9: Examples of manually captured simulated images under dynamic translation with different motion patterns.

point variation, which is sufficient for evaluating cross-view spatial reasoning.

B.4 Details of Manually Collected Data

Cross-scale perception refers to partitioning the urban environment into multiple levels of spatial granularity, for example, progressively reducing the scale from a block-level viewpoint to that of an individual building (or facility), and further down to the object level, such as billboards, trees, or sculptures. With the assistance of experienced simulator operators in our team, we collected cross-scale image sequences under large-range urban motions, resulting in a total of 246 sequences.

As shown in Fig. 9(a), we select a visually distinctive target as an anchor point, observe it from a long distance, gradually approach it, and finally localize the target for close-range observation. During this process, the number of captured images is not fixed (typically 4–8), but is instead determined by the observability of the target.

As illustrated in Fig. 9(b), we perform takeoff or landing from an open plaza with salient landmarks. Throughout this process, we record the changes in perceived urban scale, with the number of images ranging from 3 to 6.

As shown in Fig. 10, we collected 42 sets of panoramic observations at multiple locations in

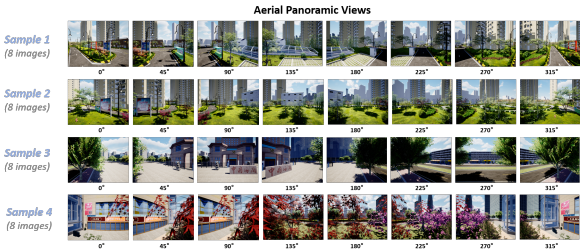


Figure 10: Examples of simulated images captured manually in rotation (panoramic views).

EmbodiedCity. Each set consists of eight camera images with a 90-degree field of view, which together form a complete 360-degree observation for the aerial agent.

B.5 Details of Multi-choice QA Generation

B.5.1 Role Templates

We adopt explicit role templates to frame the VLM as an embodied agent or an expert evaluator, which has been shown to be critical for inducing structured reasoning behaviors and consistent input–output formats in large language models.

Multi-scale Role (Fig. 11). The first role template frames the model as an observer reasoning across a hierarchy of spatial scales, ranging from regional to building and detail levels. By explicitly labeling scale transitions and enforcing scale-aware placeholders, this template encourages hierarchical spatial reasoning and object emergence analysis, which aligns with human spatial cognition theories emphasizing multi-level environmental representation.

Egocentric–Allocentric Role (Fig. 12). The second role template assigns the model the task of jointly reasoning over egocentric street-level views and allocentric aerial views of the same urban area. This design explicitly bridges self-centered perception and map-like global understanding, a distinction widely studied in cognitive science and embodied navigation. Such a role formulation enables the model to align local observations with global spatial context.

Surrounding Multi-view Role (Fig. 13). The third role template positions the VLM as an agent observing an urban object or location from multiple surrounding viewpoints, forming a 360-degree or multi-angle observation. This setting encourages cross-view consistency reasoning and spatial relation inference, which is essential for robust scene understanding under viewpoint variations.

MAIN INSTRUCTIONS – DYNAMIC SCALE

If mode == "horizontal movement":
mode_desc="The camera is moving horizontally, simulating an approach towards the target."
Else:
mode_desc = "The camera is moving vertically, simulating a descent or ascent towards the ground."

ROLE: You are given multiple images of the same urban scene captured at different scale levels. Each image will be clearly labeled with its scale identifier (e.g., Regional Scale, Building Scale, Detail Scale, etc.) immediately before the image. **{mode_desc}**
The images typically follow a progression from large scale (regional/area level) to medium scale (building level) to small scale (detail level), showing the same location with increasing zoom/ magnification.
PLACEHOLDERS:
- <obj>: Categories or names. Named as <tree1>, <building1>, <window3>, <signboard1>, etc.
- <view>: View ID representing scale level in sequence (like large→medium→small), named as <view1>, <view2>, <view3>, etc.
- <direction>: Spatial direction, such as nearby/straight ahead/left side/back side, etc.
- <scale>: Scale hierarchy, such as "Regional level / Building level / Detail level"
- <action>: Camera movement actions, such as forward translation, backward translation, arise descent, etc.
- <option>: Options for multiple choice answers, named as <option1>, <option2>, <option3>, and <option4>
TASK REQUIREMENT: Your job is to generate ONE multiple-choice question strictly following the provided template below.
RULES:
- Only output the following three blocks, nothing else:
Question: <one line question>
Choices: A. ... B. ... C. ... D. ... (and E. ... if provided)
Reasoning Process: ...
Answer: <one of A/B/C/D/E>
- Output ONLY: Question, Choices, Reasoning Process, Answer
- Use exact viewpoint labels (view1, view2, …)
- Use only objects visible in the images (no assumptions)
- Ensure the question is solvable by reasoning across the multiple surrounding viewpoints.
- Focus on Scale transition, Object emergence at different scales, Hierarchical relationships, Spatial detail changes, and Scale-aware reasoning.
- If this task is NOT APPLICABLE to the given images (e.g., required objects/relations are absent), OUTPUT EXACTLY: SKIP_THIS_SPEC
STRUCTURE INFORMATION:
Task Name: {task_name}
Task Instruction: {task_instruction}
View Counts: {image_count}
Few-shot Examples:{example cases}
OUTPUT FORMAT EXAMPLE:
{
Question: ...?
Choices: A. ... B. ... C. ... D. ... E. ...
Reasoning Process: ...
Answer: B
}

Figure 11: Prompt template for role playing (Dynamic-scale views).

MAIN INSTRUCTIONS – EGO-ALLOCENTRIC

ROLE: You are given two images of the same urban area:
- Aerial View: Allocentric aerial view (high-altitude city map-like view)
- Street View: Egocentric street view (first-person street-level view)
This task focuses on understanding the relationship between individual perspective (Egocentric) and environmental perspective (Allocentric).
PLACEHOLDERS:
- <obj>: Categories or names. Named as <brown building1>, <black tower1>, etc.
- <direction>: Absolute direction, such as North, South, West, East, etc.
- <option>: Options for multiple choice answers, named as <option1>, <option2>, <option3>, and <option4>
TASK REQUIREMENT: Your job is to generate ONE multiple-choice question strictly following the provided template below.
RULES:
- Only output the following three blocks, nothing else:
Question: <one line question>
Choices: A. ... B. ... C. ... D. ... (and E. ... if provided)
Reasoning Process: ...
Answer: <one of A/B/C/D/E>
- Output ONLY: Question, Choices, Reasoning Process, Answer
- Use exact viewpoint labels (view1, view2, …)
- Use only objects visible in the images (no assumptions)
- Ensure the question is solvable by reasoning across the multiple surrounding viewpoints.
- Focus on Appearance, Measurable indicators (height, size), Spatial relationships, Path planning.
- If this task is NOT APPLICABLE to the given images (e.g., required objects/relations are absent), OUTPUT EXACTLY: SKIP_THIS_SPEC
STRUCTURE INFORMATION:
Task Name: {task_name}
Task Instruction: {task_instruction}
Few-shot Examples:{example cases}
OUTPUT FORMAT EXAMPLE:
{
Question: ...?
Choices: A. ... B. ... C. ... D. ... E. ...
Reasoning Process: ...
Answer: B
}

Figure 12: Prompt template for role playing (Ego-allocentric views).

Rotation-based Viewpoint Role (Fig. 14). The fourth role template emphasizes rotational viewpoint changes around a fixed location, with each image annotated by precise viewing directions or camera poses. By enforcing strict viewpoint identifiers, this role promotes reasoning about self-position changes relative to a static environment, supporting rotation-aware spatial inference and mental viewpoint transformation.

B.5.2 Task Templates

As shown in Fig 15 to 18, we sequentially present the task templates for each camera dynamic, because these tasks are unique to this viewpoint mode.

B.6 Quality Control

B.6.1 Blind Filter

This filtering strategy aims to exclude questions that can be resolved solely through commonsense reasoning, without relying on explicit visual evidence. Specifically, we employ multiple VLMs (six open-source models in this study) to answer each question without providing any multi-view image inputs. If all VLMs correctly predict the answer

MAIN INSTRUCTIONS - ORBIT

ROLE: You are given multiple images captured from different viewpoints (view1, view2, …) surrounding the same urban object or location, forming a 360-degree or multi-angle observation.
The task requires reasoning about urban scenes and spatial relationships across these viewpoints.
PLACEHOLDERS:
- <obj>: Categories or names. Named as <car1>, <car2>, <brown building1>, <black tower1>, etc.
- <view>: View ID in multi-view arrangements, named as <view1>, <view2>, <view3>, etc.
- <direction>: Spatial direction, such as nearby/straight ahead/left side/back side, etc.
- <count>: Counts of objects, such as <count1>, <count2>, <count3>, etc.
- <action>: Camera movement actions, such as forward translation, backward translation, clockwise rotation, etc.
- <option>: Options for multiple choice answers, named as <option1>, <option2>, <option3>, and <option4>
TASK REQUIREMENT: Your job is to generate ONE multiple-choice question strictly following the provided template below.
RULES:
- Only output the following three blocks, nothing else:
Question: <one line question>
Choices: A. ... B. ... C. ... D. ... (and E. ... if provided)
Reasoning Process: ...
Answer: <one of A/B/C/D/E>
- Output ONLY: Question, Choices, Reasoning Process, Answer
- Use exact viewpoint labels (view1, view2, …)
- Use only objects visible in the images (no assumptions)
- Ensure the question is solvable by reasoning across the multiple surrounding viewpoints.
- Focus on Spatial relationships, C ounting, Direction understanding, Mental rotation, and Perspective transformation.
- If this task is NOT APPLICABLE to the given images (e.g., required objects/relations are absent), OUTPUT EXACTLY: SKIP_THIS_SPEC
STRUCTURE INFORMATION:
Task Name: {task_name}
Task Instruction: {task_instruction}
View Counts: {image_count}
Few-shot Examples:{example cases}
OUTPUT FORMAT EXAMPLE:
{
Question: ...?
Choices: A. ... B. ... C. ... D. ... E. ...
Reasoning Process: ...
Answer: B
}

Figure 13: Prompt template for role playing (Orbit views).

MAIN INSTRUCTIONS - ROTATION

ROLE: You are given multiple images of the same urban area taken from different viewpoints. Each image will be clearly labeled with its viewpoint identifier immediately before the image. {format_type}. The viewpoint labels provided with each image that indicate the exact viewing direction or camera position. This task focuses on understanding urban environments and the relationship between self-position and scene context through multi-view rotation.

PLACEHOLDERS:

- <obj>: Categories or names. Named as <brown building1>, <black tower1>, etc.
- <view>: View ID in multi-view arrangements, named as <view1>, <view2>, <view3>, etc.
- <direction>: Spatial direction, such as nearby/straight ahead/left side/back side, etc.
- <option>: Options for multiple choice answers, named as <option1>, <option2>, <option3>...

TASK REQUIREMENT: Your job is to generate ONE multiple-choice question strictly following the provided template below.

RULES:

- Only output the following three blocks, nothing else:
  Question: <one line question>
  Choices: A. ... B. ... C. ... D. ... (and E. ... if provided)
  Reasoning Process: ...
  Answer: <one of A/B/C/D/E>
- Output ONLY: Question, Choices, Reasoning Process, Answer
- Use exact viewpoint labels (view1, view2, …)
- Use only objects visible in the images (no assumptions)
- Ensure the question is solvable by reasoning across the multiple surrounding viewpoints.
- Focus on Spatial relationships, Object counting, Direction understanding, Mental rotation, and Perspective transformation.
- If this task is NOT APPLICABLE to the given images (e.g., required objects/relations are absent), OUTPUT EXACTLY: SKIP_THIS_SPEC

STRUCTURE INFORMATION:

Task Name: {task_name}
Task Instruction: {task_instruction}
Few-shot Examples: {example cases}

OUTPUT FORMAT EXAMPLE:

{
Question: ...?
Choices: A. ... B. ... C. ... D. ... E. ...
Reasoning Process: ...
Answer: B
}

Figure 14: Prompt template for role playing (Rotation views).

under this setting, the question is discarded, indicating that it can be solved using general perceptual priors alone and does not require detailed visual interpretation of the scene.

This procedure ensures that the retained questions genuinely demand complex image-grounded reasoning and the integration of information across multiple viewpoints, thereby sharpening the dataset's emphasis on challenging visual perception tasks. Conversely, questions for which all VLMs fail to produce the correct answer are also removed, as such cases suggest that the question may be ill-posed, ambiguous, or outside the intended scope of the task.

Together, this bidirectional filtering mechanism ensures that the final dataset consists exclusively of questions that require authentic multi-view visual reasoning—namely, those that cannot be trivially answered by all models, yet remain solvable by some VLMs when visual input is absent.

B.6.2 Details of Human Refinement

This stage adopts a structured two-stage human refinement protocol to ensure annotation reliability and consistency. We recruit volunteers on campus and pay them reasonable compensation commensurate with the region. In detail, we recruit ten annotators with master's or doctoral training and research experience related to urban spatial understanding. Annotators are divided into two independent groups with complementary roles: a *refinement group*, responsible for revising automatically generated QA pairs, and a *verification group*, responsible for independent validation and consistency checking.

Each QA instance is reviewed by at least one annotator from each group, ensuring dual human coverage. Disagreements between the two groups are explicitly recorded and resolved through adjudication by the verification group, which serves as the final decision authority. This process implicitly enforces inter-annotator consistency by requiring agreement across independent reviewers before acceptance.

During refinement, annotators explicitly flag ambiguous or ill-posed questions, including cases with unclear spatial references, underspecified viewpoints, or multiple plausible answers. Such cases are either revised through rewording and answer option correction or removed entirely if ambiguity cannot be reliably resolved. This auditing process prevents semantically underspecified samples from entering the benchmark.

To further characterize refinement outcomes, we categorize common annotation issues encountered during this stage, including (i) incorrect spatial relations, (ii) misleading or visually unsupported answer options, (iii) inconsistent reasoning chains, and (iv) viewpoint-dependent ambiguities. Accepted QA pairs undergo final proofreading to eliminate residual ambiguities and formatting errors, and human-authored reasoning processes are added to improve clarity and interpretability.

The refinement interface used by annotators is illustrated in Fig. 19.

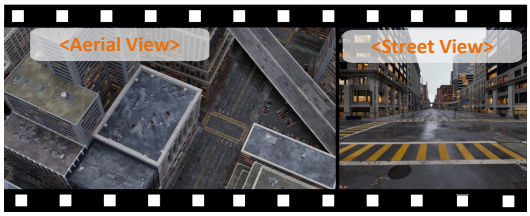
Overall, this refinement protocol prioritizes annotation reliability over scale, ensuring that subtle geometric relations are consistently grounded in the provided multi-view visual evidence.

C Experimental Details

All local model inference and fine-tuning is performed on 4×A100-SXM4-80GB. The code of our program is available at anonymous.4open.science/r/CityCube-Bench-9E72/.

C.1 Brief Introduction on Baselines

Our evaluation covers both proprietary and open-source MLLMs trained to receive multi-image inputs. The evaluated models, as well as random and



Task: Spatial Relation
Subtask: Height Estimation
Question: From the street view, estimate the height ratio between the reddish-brown building on the left and the dark building on the right.
Answer: Choices can be
A. The reddish-brown building is about 2+ times higher than the dark building.
B. The reddish-brown building is slightly higher than the dark building (1-2 times).
C. Similar height.
D. The dark building is slightly higher than the reddish-brown building.

Task: Spatial Relation
Subtask: Between Building Clusters
Question: In the aerial view, considering the area between the tall grey buildings on the left and the dark buildings on the right of the wide main street, what is this area in the street view?
Answer: Choices can be
A. Open square B. Green space C. Parking lot D. Small commercial area

Task: Mental Reconstruction
Subtask: Occlusion
Question: In the street view, what type of area is occluded by the large multi-story building on the right side of the main street, as seen from the aerial view?
Answer: Choices can be
A. Green space B. Parking lot C. Small buildings D. Road E. Water body

Task: Comprehensive Reasoning
Subtask: Shortest Route Landmarks
Question: From the aerial view, considering the shortest route from the yellow crosswalk in the foreground to the elevated highway in the distance, which landmarks does the route pass through?
Answer: Choices can be
A. The large ornate building on the left side of the main street B. The red-brick building further down the main street C. The street-level parking lot on the right side of the main street D. A and B E. B and C

Task: Comprehensive Reasoning
Subtask: Intersection Type
Question: In the aerial view, what type of traffic node is the intersection prominently visible in the foreground of the street view, marked by the yellow pedestrian crosswalks?
Answer: Choices can be
A. Major traffic hub B. Secondary intersection C. Minor junction D. Roundabout

Task: Spatial Relation
Subtask: Region Visibility
Question: What is the visibility state of the parking lot with cars, visible to the right of the main street in the aerial view, when observed from the street view?
Answer: Choices can be
A. Completely visible B. Partially visible C. Blocked by buildings D. Blocked by vegetation E. Completely invisible

Task: Mental Reconstruction
Subtask: Region Visibility
Question: What is the visibility state of the parking lot with cars, visible to the right of the main street in the aerial view, when observed from the street view?
Answer: Choices can be
A. Completely visible B. Partially visible C. Blocked by buildings D. Blocked by vegetation E. Completely invisible

Task: Comprehensive Reasoning
Subtask: Scale Understanding
Question: What is the actual scale of the prominent light grey building on the left side of the street, as observed from the street view?
Answer: Choices can be
A. Large building (>30m) B. Medium building (10-30m) C. Small building (<10m) D. Street facility

Task: Comprehensive Reasoning
Subtask: Direction to Target
Question: To reach the first side street on the left, visible in the aerial view, from the camera position in street view, which direction should be taken?
Answer: Choices can be
A. Straight ahead B. Turn left C. Turn right D. Go backward

Task: Spatial Relation
Subtask: Direction from Street Position
Question: From aerial view, identify what is in the front of your position in street view.
Answer: Choices can be
A. Red high-rise building B. White residential building C. Intersection D. Classical city hall

Task: Mental Reconstruction
Subtask: Prediction Moving
Question: Based on aerial view, predict what is most likely to be seen after moving 100 meters forward from street view.
Answer: Choices can be
A. Bridge B. Square C. Commercial area D. Residential area E. Park

Task: Perspective Taking
Subtask: Street->Aerial Region
Question: In street view, identify the pole with multiple road signs located on the left sidewalk near the yellow pedestrian crossing stripes. In aerial view, in which region is this pole located?
Answer: Choices can be
A. Northeast region B. Southeast region C. Southwest region D. Northwest region

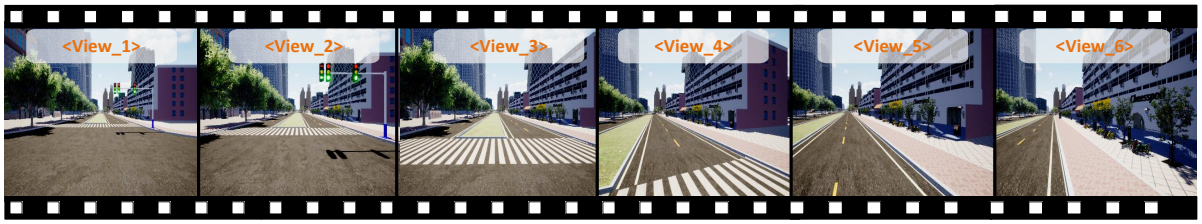
Task: World Knowledge
Subtask: Co-visibility Landmarks
Question: Which of the following landmarks appear in BOTH aerial view and street view?
Answer: Choices can be
A. Tall building with grid windows B. Ornate building on the left C. Elevated highway bridge D. Pedestrian crossing with yellow stripes E. All of the above

Task: World Knowledge
Subtask: Urban Service
Question: A traffic signal is visible on the left side of the street, at the next intersection, in the street view. How is its position marked in the aerial view?
Answer: Choices can be
A. Northeast corner of intersection B. Southwest corner of intersection C. North side of main road D. South side of main road

Task: Perspective Taking
Subtask: Aerial->Street Direction
Question: In the aerial view, locate the large building with the distinctive grey roof and columns. In which direction is this building located from the street view?
Answer: Choices can be
A. Left side B. Right side C. Straight ahead D. Behind E. Not visible

Task: World Knowledge
Subtask: Building Shape from Aerial
Question: What shape does the large ornate building on the left in the street view appear to have when viewed from the aerial view?
Answer: Choices can be
A. Cube B. C-shaped C. L-shaped D. Circular E. Other

Figure 15: Examples of task templates and their corresponding images (Ego-allocentric views).



Task: World Knowledge
Subtask: Object First Appearance
Question: In the six views from large to small scale <view1>, <view2>, ...<view6>, at which scale does the yellow signboard on the right-side building first appear?
Answer: Choices can be
A. <view1> (Regional)
B. <view3> (Building)
C. <view6> (Detail) D. Not visible

Task: World Knowledge
Subtask: Object Count Change
Question: As the scale changes from view1 (Regional Scale) through the subsequent views to view6, how does the number of observed bicycles within the frame change?
Answer: Choices can be
A. Gradually increases
B. Gradually decreases
C. Remains constant D. Uncertain

Task: World Knowledge
Subtask: Object Function
Question: What function is the yellow rectangular sign on the ground floor of the building on the right, visible in view3 (Detail Scale), most likely to provide?
Answer: Choices can be
A. Building entrance
B. Commercial advertising
C. Residential facility D. Public signage

Task: Spatial Relation
Subtask: Scale Correspondence
Question: Based on multi-scale observation, if the large, long building on the right side of the street is seen in view1 (Regional Scale), what feature appears clearer at its corresponding position in view2 (Building Scale)?
Answer: Choices can be
A. Windows B. Balconies
C. Doors D. Not visible

Task: Spatial Relation
Subtask: Spatial Relationship Change
Question: During the scale change from view1 (Regional Scale) to view6 (Scale 6), how has the spatial relationship between the Patterned Building on the right and the Distant Twin Towers changed?
Answer: Choices can be
A. Closer B. Further apart C. Remains unchanged D. Cannot determine

Task: Perspective Taking
Subtask: Reverse Scale View
Question: If you stand next to the prominent zebra crossing at view3 (Detail Scale) and look outward, what scale scene would you first see?
Answer: Choices can be
A. Back to view2 (Building scale)
B. Directly see view1 (Regional scale)
C. Still detail scale D. Not visible

Task: Mental Reconstruction
Subtask: Detail Prediction on Zoom
Question: If you are currently at view1 (Regional Scale) and gradually zoom to view2 (Building Scale), what new details would you expect to see on the building to the right?
Answer: Choices can be
A. windows B. balconies
C. signboards D. No new details

Task: Mental Reconstruction
Subtask: Zoom Center Prediction
Question: Assuming you see the building facade in view2 (Building Scale), if further zooming to view3 (Detail Scale), which part is most likely to become the central view?
Answer: Choices can be
A. Building top B. Building entrance
C. Building side wall D. Building back

Task: Mental Reconstruction
Subtask: Information Loss on Zoom Out
Question: If zooming out from view3 (Detail Scale) back to view1 (Regional Scale), what type of information would you most likely lose observation of?
Answer: Choices can be
A. Semantic categories of objects B. Structural details of buildings
C. Large-scale spatial layout
D. Urban background environment

Task: Spatial Relation
Subtask: Camera Movement Trend
Question: In the scale transition from view1 (Regional Scale) to view3 (Detail Scale), what is the camera movement trend?
Answer: Choices can be
A. From long shot to medium shot
B. From medium shot to close-up
C. From long shot directly to close-up
D. Uncertain

Task: Perspective Taking
Subtask: Scale Transition Sequence
Question: If you are observing the scene on the right side of the road, starting from view1 and approaching to view3, what is the camera transition In sequence?
Answer: Choices can be A. Regional→Detail→Building
B. Regional→Building→Detail
C. Building→Regional→Detail
D. Detail→Building→Regional

Task: Perspective Taking
Subtask: Cross-Scale Visibility
Question: If a person is at the entrance of view2, can they directly see the yellow rectangular signboard located on the right-hand building, which becomes more clearly visible in view3?
Answer: Choices can be
A. Yes, straight ahead B. Yes, to the side
C. No, blocked D. Uncertain

Task: Comprehensive Reasoning
Subtask: Object Localization Sequence
Question: From view1 to view6 observation, if you want to locate the 'Chessport' signboard, in what sequence do you need to lock on?
Answer: Choices can be
A. Region → Building → Signboard
B. Region → Signboard → Building
C. Building → Region → Signboard
D. Signboard → Building → Region

Task: Comprehensive Reasoning
Subtask: Hierarchical Structure
Question: Considering all views, what is the correct hierarchical structure of the visible scene from first to last?
Answer: Choices can be
A. Urban region > Building complex > Individual building > Local details
B. Urban region > Local details > Individual building
C. Individual building > Region > Local details
D. Local details > Region > Building complex

Task: Comprehensive Reasoning
Subtask: Camera Motion Pattern
Question: Considering the progression from view1 to view3 and assuming a drone gradually descends, which description best fits the camera motion pattern?
Answer: Choices can be
A. Zoom and descent combined, focus gradually concentrates B. Maintain horizontal position, only zoom lens
C. Horizontal movement first then zoom
D. Random switching, no pattern

Figure 16: Examples of simulated images captured manually in rotation (Orbit views).

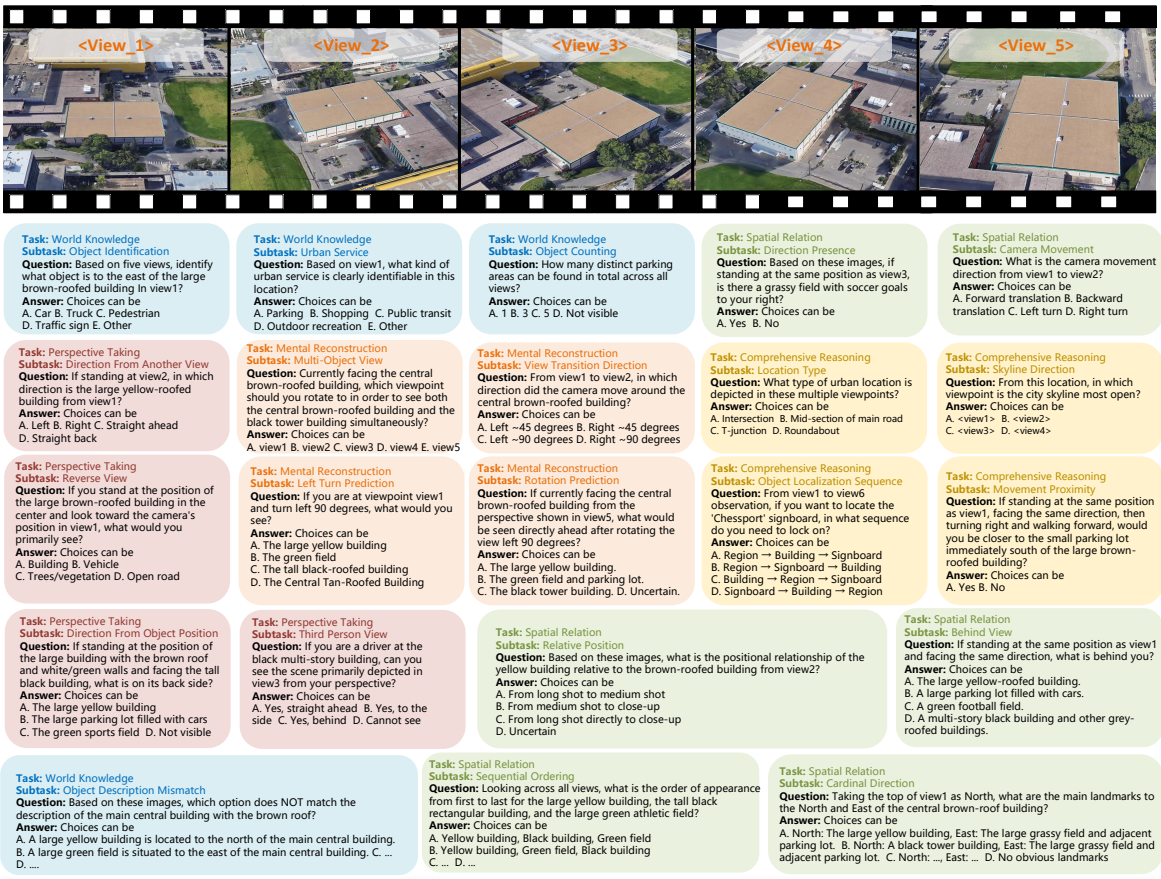


Figure 17: Examples of simulated images captured manually in rotation (panoramic views).

Figure 18: Examples of simulated images captured manually in rotation (Rotation views).

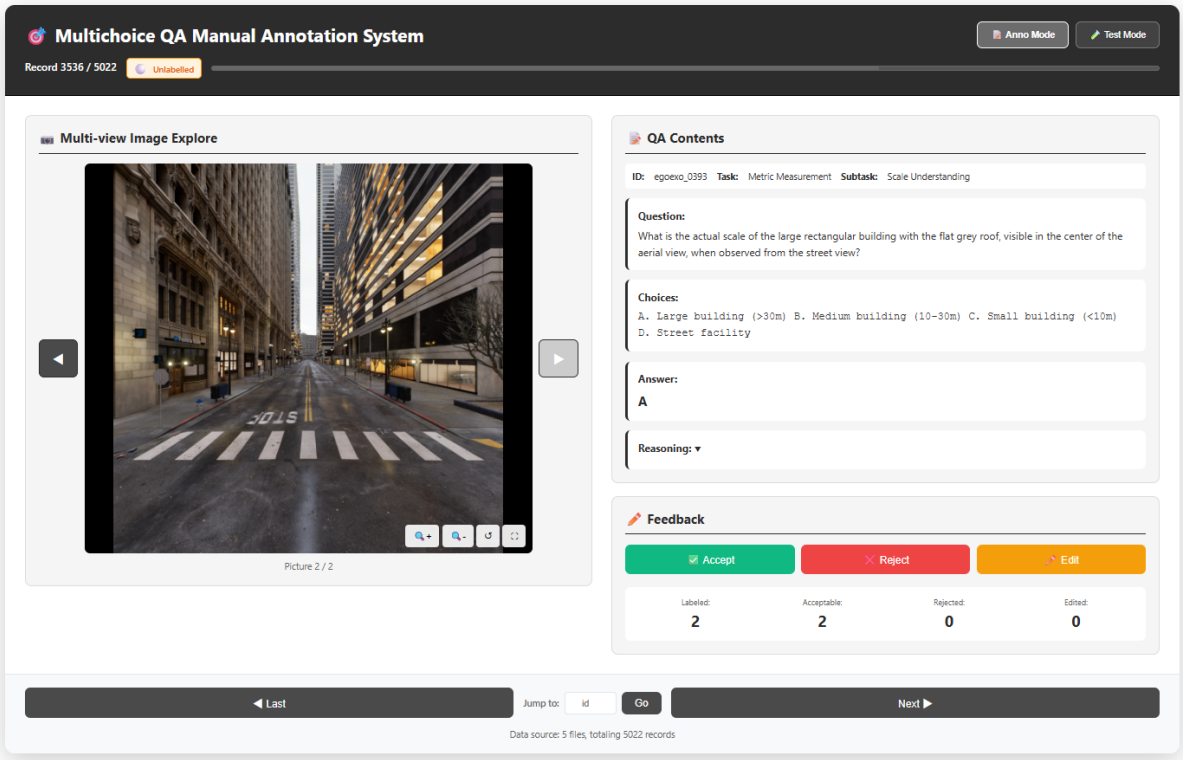


Figure 19: A demonstration Website of Multi-choice QA Annotation System on CityCube Dataset.

human baselines, are briefly introduced as follows:

Random. A random baseline model serving as the lowest performance benchmark for comparison.

Human. Human expert performance baseline representing the upper limit of human-level capability on this task.

GPT-5.1. A state-of-the-art proprietary multimodal large language model from OpenAI (OpenAI, 2025), featuring advanced visual reasoning and long-context understanding capabilities.

Gemini-2.5-Pro. A high-performance multimodal model developed by Google (Comanici et al., 2025), designed for complex reasoning over visual and textual inputs with strong generalization ability.

Qwen-3-VL-Plus. A large-scale vision–language model from Alibaba (Bai et al., 2025), optimized for multi-image understanding and instruction-following across diverse visual reasoning tasks.

Step-1o-turbo-vision. A reasoning-optimized vision-language model from StepFun, demonstrating competitive performance on multi-step and compositional visual reasoning benchmarks.

Doubao-seed1.6-251015. A proprietary multimodal model from ByteDance (Guo et al., 2025), designed to support general-purpose vision–language understanding and reasoning tasks.

Skywork-R1V4-Lite. The largest vision–language reasoning model from SkyworkAI (Wei et al., 2023), focusing on high-performance inference while maintaining strong visual comprehension capabilities.

Qwen3-VL. An open-source vision–language model family supporting multi-image inputs, widely adopted as a strong baseline for multimodal reasoning and perception tasks.

GLM-4.1V. A multimodal extension of the GLM series, featuring enhanced visual understanding and cross-modal reasoning abilities.

Kimi-VL-A3B. A compact vision–language model emphasizing efficient visual perception and instruction-following under limited parameter budgets.

MiMo-VL-7B. A 7B-parameter open-source vision–language model designed for general multimodal understanding and reasoning.

MiniCPM-V-4.5. A lightweight multimodal model from the MiniCPM series, targeting efficient deployment with competitive visual reasoning performance.

Ovis2.5. A medium-scale vision–language model from Alibaba, supporting multi-image inputs and fine-grained visual reasoning.

LLaVA-NeXT-Video. An extension of LLaVA-NeXT tailored for video and multi-frame visual reasoning, enabling temporal understanding across sequential visual inputs.

LLaVA-OneVision. A unified vision–language model supporting multiple visual modalities and tasks within a single architecture.

InternVL2.5. A strong open-source vision–language model with enhanced multi-image reasoning and long-context modeling capabilities.

Skywork-VL. A general-purpose open-source vision–language model designed for multi-modal perception and reasoning tasks.

Molmo-7B. A 7B-parameter multimodal model from AllenAI, focusing on robust visual understanding and reasoning across diverse scenarios.

Phi-4-multimodal-instruct. A compact multimodal instruction-tuned model from Microsoft, emphasizing reasoning efficiency and controllability.

Spatial-SSRL. A specialized vision–language model designed for spatial reasoning and structured scene understanding.

SpaceOm. A spatially-aware multimodal model focusing on object-level and relational reasoning in complex visual environments.

SpaceThinker. A reasoning-centric vision–language model explicitly designed to enhance spatial cognition and multi-step visual reasoning.

D Detailed Error Analysis

To provide a deeper understanding of the challenges faced by VLMs in urban spatial reasoning, we detail the specific failure cases illustrated in Fig. 20. These examples correspond to the four primary error modes discussed in the main text.

D.1 Small Object Perception Error

Case Description: As shown in the first column of Fig. 20, the model (Seed 1.6) was tasked with identifying the object located on the right side of a specific viewpoint in a street scene (View 4). The ground truth answer is "The paved area with grid patterns" visible in View 5. Failure Analysis: The model incorrectly selected "The green shelter" , explicitly stating in its reasoning that the grid-patterned paved area "is not visible". This demonstrates a failure to perceive low-salience or small-scale objects (the paved area) within the complex urban scene, leading to a hallucinated spatial arrangement where it defaulted to a more prominent object (the shelter) despite it not being the correct answer for the specific spatial query.

D.2 Spatial Reasoning Error

Case Description: In the second case, the model (Gemini 2.5 pro) was asked to determine what major landmark lies directly behind the camera's position in an aerial shot (View 1). The correct answer, based on the global context, is "A street with many small, dark-roofed houses". Failure Analysis: The model incorrectly chose "Two large, white-roofed buildings". The reasoning reveals that the model attempted to deduce the rear view by referencing other views (View 3 and 4) , but it failed to correctly orient these views relative to View 1. It stated that these views show the scene from the "back side," which was a misinterpretation of the spatial layout. This highlights the model's inability to perform accurate mental rotation or establish a coherent 360-degree spatial representation from disjointed views.

D.3 Cross-view Consistency Error

Case Description: The third column illustrates a task requiring the model (Qwen3-VL) to determine the shape of a building seen in a street view when viewed from above (Aerial view). The target building is a "Cube". Failure Analysis: The model incorrectly reasoned that the building is "L-shaped". The reasoning log shows that instead of visually grounding the specific building in the provided aerial image, the model relied on a heuristic from the street view perspective ("maybe it's a building that has a corner"). This led to a fabrication of the top-down geometry rather than a retrieval of visual evidence from the aerial modality, exemplifying a disconnection between the ground-level and overhead perspectives.

D.4 Dynamic Understanding Error

Case Description: In the final example involving motion dynamics, the model (Gemini 2.5 pro) was asked to judge how the spatial relationship (distance) between buildings on opposite sides of the street changes as the camera moves from View 1 to View 6. The ground truth is that they appear "Further apart" due to the perspective change. Failure Analysis: The model incorrectly predicted the objects would get "Closer". The reasoning indicates a fundamental error in perceiving the direction of motion; the model believed the camera was moving "backward" and that objects were receding. In

Figure 20: The four common failure cases for VLMs in multi-view spatial reasoning.

reality, the camera motion or perspective shift was misinterpreted, leading to a reversed understanding of the perspective effect.

E Implementation Details

The dataset is split into training and test sets with a ratio of 9:1, and data loading is parallelized using four worker processes. The fine-tune experiment results are produced on test set.

E.1 Training Details

We fine-tune **Qwen3-VL-2B**, **Qwen3-VL-4B** and **Qwen3-VL-8B** using supervised fine-tuning with Low-Rank Adaptation (LoRA). As described in Tab 3, the models are trained for 5 epochs using the Adam optimizer with an initial learning rate of 1×10^{-4} and a warmup ratio of 0.05. To reduce memory consumption while maintaining training stability, we employ bfloat16 precision, gradient accumulation with 4 steps, and FlashAttention. The maximum input sequence length is set to 4096 tokens, and images are resized to ensure the total pixel count does not exceed 1.6M.

LoRA is applied to all linear layers with a rank of 8 and a scaling factor of 32. We do not freeze the vision encoder or the vision-language alignment module to allow full end-to-end adaptation. We further enable padding-free training and sample packing to improve computational efficiency.

During training, evaluation and checkpoint saving are performed every 100 steps, and at most three checkpoints are retained.

Table 3: Training hyperparameters for **CityBot** series.

Category	Hyperparameter	Value
Training Setup	Training paradigm	Supervised fine-tuning (LoRA)
	Number of epochs	5
	Precision	bfloat16
Batching	Per-device train batch size	1
	Per-device eval batch size	1
	Gradient accumulation steps	4
	Effective batch size	4
Optimization	Learning rate	1×10^{-4}
	Warmup ratio	0.05
LoRA Configuration	LoRA rank (r)	8
	LoRA scaling factor (α)	32
	Target modules	All linear layers
	Training scope	Full model (no freezing)
Architecture & Memory	Attention implementation	FlashAttention
	Padding-free training	Enabled
	Sample packing	Enabled
	Gradient checkpointing	Enabled (ViT excluded)
Input	Maximum sequence length	4096
	Maximum image pixels	1,605,632
Evaluation & Logging	Evaluation interval	Every 100 steps
	Checkpoint saving interval	Every 100 steps
	Max checkpoints kept	3
	Logging interval	Every 10 steps
Data Loading	Test split ratio	0.1
	Dataset preprocessing workers	4
	Dataloader workers	4

E.2 Human Evaluation Details

As shown in Figure 21, we collected a set of volunteer evaluation results through an interactive web interface. These volunteers are recruited independently and unrelated to the annotators who participated in refining the questions. They also did not receive any extra training in spatial knowledge.

F Further Discussion

3D Question Answering A precise understanding of objects and their spatial relationships within 3D scenes is critical for applications in robotics and autonomous driving. Leveraging the holistic 3D scene datasets such as ScanNet (Dai et al., 2017) and Matterport3D (Chang et al., 2017), the research community has developed a variety of 3D Question Answering (3DQA) benchmarks like ScanQA (Azuma et al., 2022; Achlioptas et al., 2020; Hong et al., 2023). The core paradigm of these benchmarks involves inferring a comprehensive 3D understanding from a limited set of 2D views. In line with this, employing multiple views as input for the LLMs has become a primary methodology in 3DQA (Fu et al., 2024; Huang et al., 2024; Guo et al., 2023). Our work extends this research trajectory and systematically incorporates typical view combinations found in existing literature and, more importantly, pioneers the extension of the 3DQA application domain from indoor environments to the more expansive and challenging context of urban spaces.

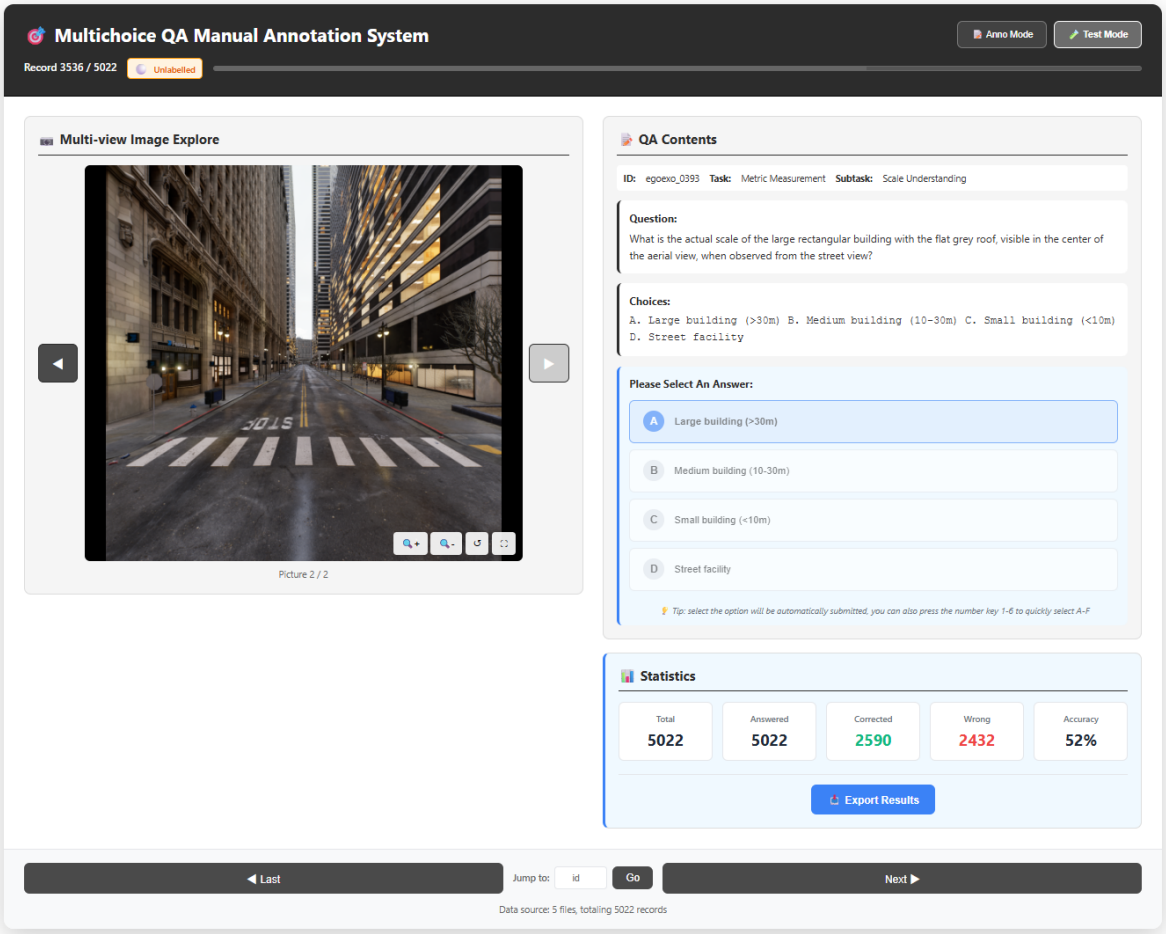


Figure 21: A demonstration Website of Multi-choice QA Test System of CityCube Dataset.