RoadBench: Benchmarking MLLMs on Fine-Grained Spatial Understanding and Reasoning under Urban Road Scenarios

Jun Zhang[1] *, Xin Zhang[1] *, Jie Feng[2] †, Long Chen[3], Junhui Wang[3], Zhicheng Liu[3], Depeng Jin[1], and Yong Li[1] †

[1] Department of Electronic Engineering, BNRist, Tsinghua University, Beijing, China
[2] Zhongguancun Academy, Beijing, China
[3] Amap, Alibaba Group, Beijing, China
fengjie@bza.edu.cn, liyong07@tsinghua.edu.cn

Abstract. Multimodal large language models (MLLMs) have demonstrated powerful capabilities in general spatial understanding and reasoning. However, their fine-grained spatial understanding and reasoning capabilities in complex urban scenarios have not received significant attention in the fields of both research and industry. To fill this gap, we focus primarily on road markings as a typical example of fine-grained spatial elements under urban scenarios, given the essential role of the integrated road traffic network they form within cities. Around road markings and urban traffic systems, we propose **RoadBench**, a systematic benchmark that comprehensively evaluates MLLMs' fine-grained spatial understanding and reasoning capabilities using Bird's-Eye View (BEV) and First-Person View (FPV) image inputs. This benchmark comprises eight tasks consisting of 3,040 strictly manually verified test cases, constructed from 2,137 unique BEV images and 721 unique FPV images collected from five Chinese cities with relatively consistent traffic conventions. These tasks form a systematic evaluation framework that bridges understanding at local spatial scopes to global reasoning. They not only test MLLMs' capabilities in recognition, joint understanding, and reasoning but also assess their ability to integrate image information with domain knowledge. After evaluating 20 mainstream MLLMs, we confirm that RoadBench is a challenging benchmark for MLLMs while revealing significant shortcomings in existing MLLMs' fine-grained spatial understanding and reasoning capabilities within urban scenarios. In certain tasks, their performance even falls short of simple rule-based or random selection baselines. These findings, along with RoadBench itself, will contribute to the comprehensive advancement of spatial understanding capabilities for MLLMs. The benchmark code is available at https://github.com/tsinghua-fib-lab/RoadBench, and the supplementary material provides example data, prompts, evaluation scripts, and raw evaluation results.

* Equal contribution.
† Corresponding authors.

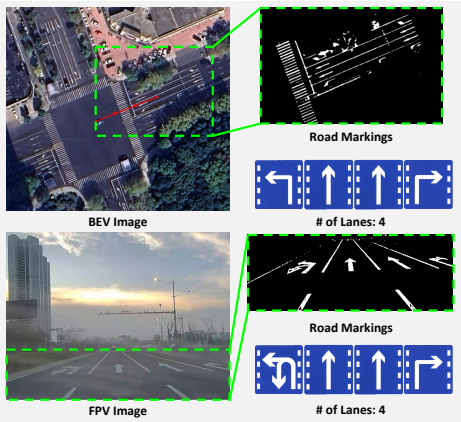

Fig. 1: Examples of road markings in BEV and FPV images. Road markings such as lane dividers and turning arrows provide fine-grained visual cues for lane counting and lane designation recognition.

1 Introduction

Multimodal large language models (MLLMs) have become a crucial tool for recognizing and understanding the real world due to their powerful combined visual-language comprehension and reasoning capabilities [1, 4, 32, 43]. They are progressively replacing specialized models in fields such as satellite image recognition [16, 46, 47], autonomous driving [11, 34, 48], embodied intelligence [13, 23], etc. Spatial understanding and reasoning in urban environments, as one of the key real-world application scenarios for MLLMs, have also garnered significant attention recently [15, 16, 31]. Various research efforts have released benchmarks [17, 40, 51] for spatial understanding and reasoning tasks in urban environments, evaluating MLLM performance from multiple perspectives. Existing urban spatial understanding benchmarks provide the foundational data and evaluation criteria necessary to advance MLLM research for real-world applications.

However, we observe that existing benchmarks on urban scenarios mainly focus on whole-image-level understanding or isolated object recognition of specific types, such as GeoQA, landmark recognition, and vehicle detection. There appears to be a lack of attention to understanding and reasoning about fine-grained spatial elements within urban scenarios. Recently, cutting-edge research has been exploring how to enhance MLLM's fine-grained understanding and reasoning capabilities in scenarios such as mathematical problems [45] and remote sensing images [27]. Nevertheless, studies on MLLM's such capabilities have yet to be applied to urban scenarios, particularly lacking evaluation benchmarks tailored to real-world urban settings.

Under urban scenarios, a representative example of such fine-grained spatial structural elements is the markings painted on city road surfaces shown in Figure 1. These narrow and long lines and arrows, drawn based on specialized

traffic knowledge, collectively form an integrated system that effectively divides and organizes urban space, thereby regulating the movement of pedestrians and vehicles. The tasks of understanding and reasoning about road markings pose significant challenges to MLLM's capabilities:

- The ability to recognize fine-grained structures at a global scale, as road markings are typically thin and extend across the entire image.
- The joint understanding and reasoning of multiple fine-grained structures, as road markings require overall consideration for accurate semantic recognition.
- The integration of image information and domain knowledge to generate reasonable responses.

We believe that developing MLLM's fine-grained understanding and reasoning capabilities under urban scenarios will not only advance its applications in transportation fields, such as high-definition map auto-generation and end-to-end autonomous driving. Advancing fine-grained urban road reasoning can also strengthen the general visual-spatial reasoning abilities of MLLMs, as it requires models to understand complex spatial layouts, directional relations, and structured scene semantics.

Therefore, we propose **RoadBench**, a systematic benchmark primarily designed to evaluate MLLM's understanding and reasoning capabilities regarding fine-grained spatial structural elements under urban scenarios. RoadBench provides a rich collection of annotated Bird's-Eye View (BEV) and First-Person View (FPV) images, comprising a total of 3,040 test cases across 8 tasks. These tasks include not only lane counting and lane designation recognition, which are directly based on understanding and reasoning about road markings, but also extend to road network correction and road type classification tasks that rely on joint reasoning involving both road markings and other information within the image. The test cases comprising RoadBench originate from manually selected images across multiple major cities. All labels and ground truth data have been labeled and checked by humans to ensure accuracy, with all potentially privacy-sensitive information completely masked and anonymized. These test cases encompass diverse factors that may impact MLLM understanding, including road and intersection patterns, lighting conditions, seasons, and image resolution. Such diversity supports the systematic evaluation of MLLMs' fine-grained spatial understanding and reasoning capabilities. Based on RoadBench, we conducted a systematic evaluation of 20 mainstream closed-source and open-source MLLMs. Benchmark results indicate that MLLMs failed to achieve satisfactory outcomes across various tasks, even underperforming against baselines based on random choice or simple rules in certain tasks. This highlights both the current limitations of MLLMs in achieving fine-grained spatial understanding and reasoning capabilities within urban environments, and underscores RoadBench as a highly challenging benchmark for MLLMs to evaluate such capabilities.

Overall, the main contributions of this paper include the following points:

- We propose a systematic benchmark named **RoadBench**, designing eight tasks under both BEV and FPV image inputs to comprehensively evaluate

MLLMs' understanding and reasoning capabilities regarding fine-grained spatial elements under urban scenarios from local to global scopes.
- We collected, processed, and annotated 3,040 test cases as datasets for the eight tasks in RoadBench using data sources such as satellite imagery, online map service providers, and crowd-sourced in-vehicle camera photo databases. All test cases underwent rigorous manual verification and annotation to ensure the accuracy of their labels.
- We conducted a systematic evaluation of 20 mainstream MLLMs using RoadBench. The results demonstrate that RoadBench is a highly challenging benchmark for MLLMs, while also revealing the limitations of existing MLLMs in fine-grained spatial understanding and reasoning capabilities within urban scenarios.

2 Related Work

2.1 MLLM for Spatial Intelligence

Inspired by the powerful comprehension and complex reasoning capabilities emerging from large language models [5, 35], researchers have achieved alignment and fusion between text and image modalities through techniques such as CLIP [29] and BLIP [22] to construct multimodal large language models (MLLMs) [1, 4, 32, 43]. These MLLMs can simultaneously process textual and visual inputs to perform comprehension and reasoning tasks, enabling them to accomplish a wide range of complex operations [11, 39]. Additionally, some recent researchers have focused on enhancing MLLM's fine-grained spatial understanding and reasoning capabilities in scenarios such as mathematical problems [25, 38, 45], daily images [2, 7, 9, 19], abstract visual puzzles [30, 36], and remote sensing images [20, 27, 41]. Leveraging the capabilities of MLLM foundation models, some researchers have successfully developed MLLMs specifically tailored for urban scenarios, such as UrbanLLaVA [15, 16], UrbanMLLM [47], Embodied-R [49], VLDrive [44] and Spatial-LLM [8]. However, research on the fine-grained understanding and reasoning capabilities of MLLM in urban scenarios remains relatively rare at present. The lack of benchmarks may be one of the primary factors constraining such work. Table 1 compares RoadBench with existing urban-scene and spatial-reasoning benchmarks along key dimensions, including input view, task type, spatial granularity, and road-scene reasoning focus.

2.2 Spatial Benchmarks for MLLM

In recent years, researchers have released extensive benchmarks for MLLMs. For example, MME [18], Seed-bench [21], and MMBench [24] introduce various perception, cognition, and reasoning tasks to comprehensively evaluate the performance of MLLMs. Some recent researchers have also developed benchmarks using abstract visual puzzles to evaluate the spatial reasoning capabilities of MLLMs, for example, SPACE [30], SpatialEval [36], and Spatial457 [37]. VSI-Bench [42]

Table 1: Comparison of spatial benchmarks for MLLM under real-world urban scenarios.

Benchmark	Minimum Spatial Granularity	Element Inter-dependency	Viewpoint
CityBench [17]	★ (buildings)	★ (isolated elements)	BEV + FPV
UrBench [51]	★★ (traffic signs)	★★ (cross-view matching)	BEV + FPV
CityEQA [50]	★★ (cars & buildings)	★ (isolated elements)	3D Scene (synthetic)
DriveBench [40]	★★ (cars)	★ (isolated elements)	FPV
NuPlanQA [28]	★★ (cars & traffic signs)	★★ (trajectory reasoning)	BEV + FPV
MapDR [6]	★★ (lanes & traffic signs)	★★ (rule-lane matching)	FPV
RoadBench	**★★★ (road markings)**	**★★★ (topological + regulatory logic)**	**BEV + FPV**

and MM-Spatial [12] evaluate the spatial understanding and reasoning abilities of MLLMs with indoor environment. In real-world urban scenarios, CityBench [17] primarily evaluates MLLM's ability to perceive and understand images about cities, as well as its planning and decision-making capabilities. UrBench [51] focuses on evaluating MLLMs in cross-view urban scenarios. CityEQA [50] investigates the urban scenarios from the aerial vehicle in a realistic 3D urban simulator. DriveBench [40] and NuPlanQA [28] focus on evaluating the reliability of MLLM and multi-view understanding abilities in autonomous driving applications. MapDR [6] is a benchmark dataset designed to evaluate the ability of MLLMs to understand and interpret driving rules through traffic signs. Overall, as shown in Table 1, there is currently no benchmark attentive to the recognition, understanding, or reasoning of fine-grained spatial elements with strong interdependency like road markings under urban scenarios, which limits the evaluation of MLLM capabilities in real world scenarios.

3 RoadBench

3.1 Benchmark Overview

RoadBench is designed as a benchmark to evaluate the fine-grained spatial understanding and reasoning capabilities of MLLM models in urban scenarios. As shown in Figure 2, RoadBench contains two types of urban scene images: bird's eye view images derived from satellite imagery and first-person view images captured by in-vehicle cameras. Based on these two categories of images, we design eight benchmark tasks centered around road markings, which are common fine-grained spatial elements in urban scenarios. These tasks are organized by the spatial scope required for completion, ranging from local perception to global contextual reasoning. Furthermore, we introduce cross-view tasks that require joint reasoning across BEV and FPV perspectives to evaluate cross-view understanding. Together, they constitute a hierarchical and systematic benchmark spanning fine-grained perception, cross-view synthesis, and global reasoning. The number of test cases included in each task is also shown in Figure 2. The entire benchmark comprises a total of 3,040 test cases. These rigorously hand-curated and processed test cases cover diverse scenarios involving different road and intersection patterns, lighting conditions, seasons, and image resolutions, providing comprehensive coverage of the diversity in urban environments.

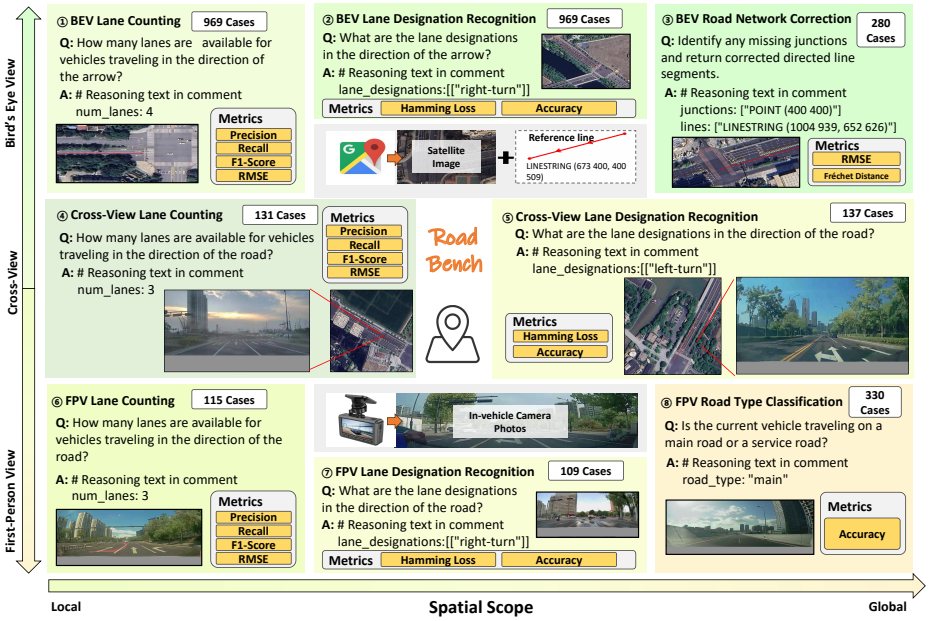

Fig. 2: The overview of RoadBench.

3.2 Benchmark Tasks

RoadBench comprises eight benchmark tasks, including lane counting and lane designation recognition (utilizing both BEV and FPV imagery), road network correction (BEV-based), and road type classification (FPV-based). Additionally, we include cross-view versions of lane counting and designation recognition to evaluate joint reasoning across perspectives. The following sections detail the setup and evaluation protocols for each task; specific MLLM prompts, representative inputs, and model responses are provided in Appendix 6.

BEV Lane Counting. The lane counting task based on BEV images requires the MLLM to determine the number of lanes contained within the road indicated by the reference line, using the input satellite image and the additional directed reference line. As shown in the example in Appendix 6.1, reference lines are drawn as polylines with prominent red arrows on satellite imagery to serve as visual prompts. The MLLM's output is required to be a text describing the reasoning process along with the number of lanes. This task aims to conduct a preliminary evaluation of MLLM's ability to follow both visual and textual prompts simultaneously while analyzing and understanding the narrow road markings surrounding reference lines in satellite imagery. The coordinates of reference lines and actual lane counts in test cases originate from the databases of a tier-1 online map service provider, which also supplies this data for online service delivery. In terms of evaluation metrics, we employed multi-class classification

metrics (Precision, Recall, F1-score) on one hand, while also utilizing RMSE to assess the deviation between MLLM outputs and ground truth.

BEV Lane Designation Recognition. The lane designation recognition task based on BEV images shares the same MLLM input setup as the BEV lane counting task, including satellite images and reference lines. In difference, this task further requires MLLM to infer the direction type of each lane by identifying and understanding road surface arrows, among other methods, when the correct number of lanes is known. Lane direction types include U-turn, left turn, straight ahead, right turn, and their combinations. This task not only challenges the MLLM's ability to jointly understand a complete set of road markings, but also tests its domain expertise in lane designation. Ground truth data also comes from the online service database. Given the combinatorial nature of lane directions, we treat the classification of each lane as an independent multi-label classification problem. We employ the Hamming Loss as the primary metric and utilize accuracy as a more stringent metric to evaluate whether MLLMs fully comprehend the lane directions.

BEV Road Network Correction. The road network correction task based on BEV images is both highly challenging and of significant practical value. This task begins by feeding an inaccurate reference line into MLLMs. These reference lines often contain errors such as missing actual junctions, which can lead to map services providing incorrect directions to users. MLLMs are tasked with understanding the concepts of junctions and road segments, then inferring the correct junction locations and road segments based on the diverse information contained within the images to achieve road network correction. In this task, MLLMs have to recognize not only the image content within the reference line area but also understand broader contextual information such as vehicle orientation and building layout to determine whether any junctions are missing. The reference lines for this task are sourced from OpenStreetMap [†], and the actual junction and road segment data also originate from the online map service database. The evaluation of results is divided into two parts: the accuracy of junction points and the accuracy of road segment polylines. Since the number of points and polylines returned by MLLMs may not match the ground truth, both evaluations will first match the ground truth with the MLLM output based on the nearest-neighbor principle. Points or polylines that fail to match will be considered as mapped to infinity. Point matching employs Euclidean distance, and RMSE with a distance upper bound threshold is used as the evaluation metric. For polylines, given the critical role of direction in road network correction, we employ the Fréchet Distance [14], which evaluates directed polyline similarity, as both a distance criterion and a performance metric. This metric also incorporates a distance upper bound. Additionally, since image sizes vary, all coordinates are normalized to the range $[0, 1]$ based on the image length and width before entering the indicator calculation.

FPV Lane Counting. Similar to the BEV lane counting task, the lane counting task based on FPV images also requires the MLLM to determine the

[†] https://www.openstreetmap.org/

number of lanes from images, sharing identical evaluation metrics and ground truth data sources. The difference lies in replacing the image perspective with a first-person view captured by an in-vehicle camera. Simultaneously, the reference lines indicating the target road on the BEV image are removed, requiring MLLMs to understand the spatial relationship between the camera and the surrounding roads to make determinations.

FPV Lane Designation Recognition. The lane designation recognition task based on FPV images shares identical metrics and ground truth data sources with BEV lane designation recognition task. Compared to the BEV perspective, which relies entirely on road markings to identify lane direction, MLLMs in the FPV perspective can determine lane direction both through road markings and by confirming overhead signs. However, FPV images also introduce new challenges for MLLMs. Scenes captured by FPV cameras may include low-light conditions at night or situations where congested traffic obscures road markings. The real-world scenarios reflected in this data will test MLLMs' ability to synthesize information from multiple sources and produce accurate reasoning.

FPV Road Type Classification. The road type classification task based on FPV images requires MLLMs to determine whether the current vehicle is traveling on the main road or the service road based on the image content. Unlike other tasks, this task demands that MLLMs go beyond understanding and reasoning about a single category of urban spatial elements, such as road markings or even road networks. Instead, it requires MLLMs to grasp the semantic information implied by the surrounding environment as a whole and to make inferences based on common sense. For example, MLLMs can determine whether a road is a service road based on pedestrians or street-side shops in an image, or identify a main road based on enclosed fences and median strips. Since the dataset construction ensures an equal distribution of test cases between the main and service roads, this constitutes a balanced binary classification task. Thus, the evaluation metric solely employs accuracy.

Cross-View Lane Counting. Unlike single-view tasks, Cross-View Lane Counting provides geographically paired BEV and FPV imagery at the scene level, requiring the MLLM to perform joint reasoning across disparate perspectives to determine the lane count of a target road. This task exploits the complementary strengths of each modality: the BEV perspective provides a clear topological overview of road geometry, while the FPV perspective offers high-fidelity visual cues, such as lane markings and local traffic context.

To succeed, the model must establish an explicit spatial correspondence between the two views, reconciling top-down structural data with ego-centric visual features. By maintaining the same evaluation metrics and ground-truth sources as the single-view counting tasks, this benchmark provides a controlled and rigorous assessment of the model's capacity for multi-view information fusion.

Cross-View Lane Designation Recognition. Building on the cross-view framework, this task requires the MLLM to determine the specific directional designations for each lane. While the BEV perspective defines the underlying structural topology, the FPV perspective contributes essential semantic cues, such

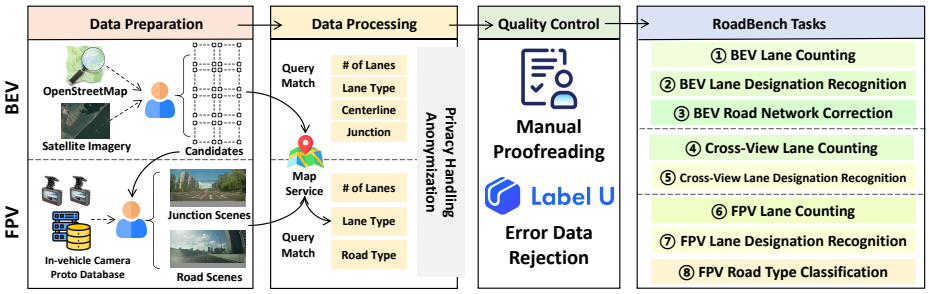

Fig. 3: The RoadBench curation pipelines to construct the datasets for the eight benchmark tasks.

as overhead signs and pavement arrows. Success hinges on the model's ability to map these ego-centric visual features onto the global road layout. Consequently, this task serves as a rigorous test of the model's cross-view semantic alignment and reasoning depth in complex, multi-lane environments.

In summary, the aforementioned eight benchmark tasks systematically evaluate MLLM's ability to understand and reason about fine-grained spatial elements in urban environments across spatial scales from local to global based on images captured from diverse perspectives and scenes.

3.3 Benchmark Curation

The curation process for RoadBench can be divided into three stages as shown in Figure 3: data preparation, data processing, and quality control. To fully ensure data quality, most steps rely primarily on human expert annotation and rule-driven programmatic automatic matching.

Data Preparation. For tasks related to BEV images, the primary step in data preparation is determining appropriate spatial scopes based on satellite image resolution and junction patterns. Specifically, we first download OpenStreetMap data from areas with relatively high satellite image resolution and extract all junctions. We then manually review the corresponding satellite imagery from Google Maps [‡] for each junction to eliminate invalid images that are fake, difficult to identify, or severely obscured, ultimately obtaining the final candidate bounding boxes. For FPV imagery, we extract valid instances from captured in-vehicle camera photos in the same junction candidates collected for the BEV tasks and road scenes from the same cities to ensure geographic consistency. For cross-view tasks, BEV-FPV pairs are constructed according to geographic proximity and road-segment correspondence, with mismatched samples manually filtered to ensure reasonable scene-level consistency. For junction scene images used in lane counting and lane designation recognition, annotators are instructed to include a certain proportion of challenging scenarios, such as nighttime conditions or

[‡] https://www.google.com/maps

obscured lane markings, to test the capabilities of MLLMs. For road scenes, the number of images for main roads and service roads keep consistent to avoid class imbalance issues.

Data Processing. After completing data preparation, the bounding box data is fed into the database of the tier-1 online map service provider to extract labels such as actual road centerlines, junction locations, the number of lanes, and lane directions within the specified area. These high-quality datasets are directly applicable to lane-related tasks as ground truth. Meanwhile, when matched with OpenStreetMap data, they form inaccurate reference lines and ground truth pairs suitable for road network correction tasks. For FPV images captured at junction and road scenes, the required labels including lane count, lane direction type, and road type are matched from the map provider's database by using the image capture coordinates. All datasets have undergone anonymization and privacy handling. All IDs have been randomized into Universally Unique Identifiers (UUIDs). All location coordinates have been processed into pixel coordinates on the images. All content in FPV images that could potentially expose the privacy of the photographer or the surrounding environment has been manually obscured using gray masks.

Quality Control. While the data construction pipeline utilizes automated matching, minor annotation errors may persist due to sensor noise, coordinate misalignments, or algorithmic edge cases. To ensure the highest level of dataset integrity, we conducted a rigorous manual audit of the final test set. Five domain experts performed a comprehensive review and refinement of all samples using the LabelU platform [26]. Notably, samples identified as ambiguous or fundamentally erroneous were systematically excluded rather than re-annotated; this conservative filtering strategy ensures the benchmark remains free from subjective human bias and maintains high signal-to-noise quality.

4 Experiments

4.1 Evaluation Settings

Evaluated MLLMs. To comprehensively evaluate the performance of MLLMs on RoadBench, we selected mainstream MLLMs released within the last one year and included both open-source and closed-source models with varying parameter counts from different providers. In the open-source models, we selected the 11B and 90B versions of LLaMA-3.2-Vision, the 7B, 32B, and 72B versions of Qwen2.5-VL [4], the 2B, 4B, 8B, 30B, 32B and 235B versions of Qwen3-VL [3], the 12B and 27B versions of Gemma-3 [32], and GLM-4.5V [33]. In the closed-source model selection, we chose Gemini-2.5-Flash, Gemini-2.5-Flash-Image, Gemini-2.5-Pro from Google [10], as well as GPT-5-Nano, GPT-5-Mini, and GPT-5 from OpenAI. Due to temporary model deprecations or API unavailability during the evaluation period, results for certain models are left blank. We also fine-tuned the Qwen3-VL models with 2B, 4B, and 8B parameters on the RoadBench training set.

Baselines. To aid in understanding the practical applicability of MLLM across these benchmark tasks, we include several rule-based or simple random

Table 2: The comprehensive evaluation results on RoadBench. The best-performing result is indicated in **bold**, and the second-best one is indicated with an underline. In the table, F1, HL, and FD are abbreviations for F1-Score, Hamming Loss, and Fréchet Distance, respectively. For BEV road network correction tasks, the normalized distance upper threshold for the RMSE metric used to evaluate junction point accuracy is 20%, while the one for the Fréchet Distance metric used to evaluate segment polyline accuracy is 50%.

Model	BEV Lane Counting		BEV Lane Designation Recognition		BEV Road Network Correction		FPV Lane Counting		FPV Lane Designation Recognition		FPV Road Type Classification	Cross-View Lane Counting		Cross-View Lane Designation Recognition	
	F1 ↑	RMSE ↓	HL ↓	Acc. ↑	RMSE ↓	FD ↓	F1 ↑	RMSE ↓	HL ↓	Acc. ↑	Acc. ↑	F1 ↑	RMSE ↓	HL ↓	Acc. ↑
LLaMA-3.2-11B-Vision	0.238	1.586	0.356	0.134	0.173	0.377	0.244	1.498	0.324	0.179	0.527	-	-	-	-
LLaMA-3.2-90B-Vision	0.282	1.379	0.214	0.457	0.148	0.306	0.309	1.165	0.166	0.514	0.655	-	-	-	-
Qwen2.5-VL-7B	0.240	1.416	0.297	0.111	0.170	0.399	0.203	1.555	0.278	0.158	0.515	0.227	1.511	0.273	0.203
Qwen2.5-VL-32B	0.254	1.255	0.239	0.361	0.155	0.314	0.307	1.099	0.236	0.310	0.542	0.226	1.248	0.206	0.408
Qwen2.5-VL-72B	0.200	1.355	0.213	0.463	0.139	0.268	0.318	1.265	0.188	0.471	0.585	0.239	1.384	0.183	0.457
Qwen3-VL-2B	0.186	1.850	0.313	0.186	0.154	0.369	0.210	1.704	0.286	0.064	0.539	0.152	1.687	0.268	0.100
Qwen3-VL-4B	0.113	2.076	0.323	0.194	0.152	0.332	0.382	1.205	0.286	0.064	0.506	0.329	1.143	0.268	0.100
Qwen3-VL-8B	0.243	1.507	0.257	0.372	0.151	0.294	0.266	1.322	0.208	0.441	0.539	0.266	1.182	0.217	0.432
Qwen3-VL-32B	0.276	1.299	0.188	0.515	0.145	0.257	0.340	1.115	0.144	0.587	0.533	0.298	1.185	0.175	0.514
Qwen3-VL-30B-A3B	0.263	1.319	0.272	0.355	0.148	0.302	0.287	1.209	0.224	0.429	0.491	0.279	1.313	0.223	0.408
Qwen3-VL-235B-A22B	0.256	1.328	0.191	0.496	0.139	0.249	0.369	1.034	0.159	0.541	0.530	0.298	1.316	0.178	0.479
Gemma-3-12B	0.207	1.214	0.256	0.341	0.156	0.289	0.200	1.212	0.263	0.343	0.548	0.196	1.201	0.261	0.347
Gemma-3-27B	0.279	1.334	0.214	0.399	0.146	**0.232**	0.250	1.285	0.195	0.404	0.512	0.248	1.139	0.224	0.381
Gemini-2.5-Flash	0.245	1.371	0.202	0.471	0.148	0.278	0.327	1.216	0.176	0.480	0.536	0.234	1.182	0.282	0.094
Gemini-2.5-Flash-Image	0.225	1.340	0.186	0.490	0.147	0.262	0.225	1.268	0.173	0.471	0.555	-	-	-	-
Gemini-2.5-Pro	0.302	1.252	0.172	0.524	0.140	0.289	**0.522**	**0.802**	0.160	0.547	**0.748**	0.365	1.095	0.282	0.094
GLM-4.5V	0.289	1.183	0.175	0.535	0.171	0.268	0.322	1.071	0.145	0.571	0.570	0.329	1.339	0.142	0.557
GPT-5-Nano	0.197	1.529	0.171	0.540	**0.132**	0.258	0.211	1.345	0.151	0.550	0.609	0.244	1.384	0.153	0.526
GPT-5-Mini	**0.355**	**1.162**	**0.103**	**0.662**	0.141	0.235	0.362	1.009	0.137	0.584	0.588	0.344	1.011	**0.121**	0.588
GPT-5	0.305	1.350	0.159	0.579	0.143	0.261	0.498	0.923	0.122	**0.626**	0.685	**0.400**	**0.878**	0.124	**0.597**
Rule-based	0.248	1.383	0.149	0.587	0.155	0.379	0.284	1.430	**0.127**	0.605	0.458	0.252	1.395	0.130	0.588
Qwen3-VL-2B(tuned)	0.319	1.204	0.232	0.338	0.114	0.190	0.387	0.987	0.103	0.678	0.830	0.421	0.957	0.105	0.668
Qwen3-VL-4B(tuned)	0.339	1.142	0.230	0.338	0.107	0.174	0.517	0.870	0.100	0.693	0.800	0.535	0.829	0.096	0.688
Qwen3-VL-8B(tuned)	0.339	1.142	0.230	0.338	0.097	0.154	0.433	0.909	0.094	0.708	0.830	0.540	0.824	0.085	0.728

baselines. In the two lane counting task, we provided two baselines: always choose two lanes and randomly select two lanes from $\{2, 3, 4\}$ with a uniform distribution. For the two lane designation recognition task, we designed a mapping table based on traffic common sense as a baseline. The details can be found in Appendix 7.2. The FPV road type classification task uses uniformly distributed random selection as the baseline. Due to the complexity of the BEV road network correction task, we directly treat the input reference line as the output road segment polyline, using the start and end points of the reference line as the coordinates of the identified intersection. The main result shown in Table 2 will only include the optimal baseline. Complete experimental results can be found in Appendix 7. In addition, since the Quality Control phase only retained data that could be correctly identified by a human, we believe that a human with the related knowledge can perform all tasks completely correctly. Therefore, no human baseline is provided for comparison.

Evaluation Metrics. Throughout the entire experiment, the performance of MLLMs across various tasks was comprehensively evaluated using the metrics shown in Figure 2. The metrics (Precision, Recall, and F1-Score) used to evaluate multi-classification tasks are weighted according to the sample size. In the BEV road network correction task, the normalized distance upper bound thresholds for RMSE and Fréchet Distance were set to $\{10\%, 20\%, 50\%\}$ to ultimately select the

appropriate threshold. Due to page size limitations, the main results in Table 2 report no more than two metrics that best reflect the performance of MLLMs. The complete results can be found in Appendix 7.

Error Handling. To minimize the interference of detectable errors on MLLMs' fine-grained spatial understanding and reasoning evaluation, the benchmark procedure incorporates a series of error-handling mechanisms. The program detects issues such as failed API calls, empty response values, and incorrect return formats, and re-invokes MLLMs with identical inputs. This retry process is limited to a maximum of six attempts. If MLLMs still fail to produce correct results after retries, the outcome is recorded as zero or empty.

4.2 Main Results

The comprehensive evaluation results on RoadBench for the selected MLLMs and baselines are listed in Table 2. By analyzing these results, we can find the following conclusions.

RoadBench is a highly challenging benchmark for MLLMs. Overall, neither the most powerful closed-source models nor open-source models perform sufficiently well on the tasks proposed by RoadBench for evaluating the fine-grained understanding and reasoning capabilities of MLLMs in urban scenarios. For example, in the BEV lane counting task, the best model GPT-5-Mini achieved an F1-Score of only 0.355, and the RMSE between the predicted lane counts and the ground truth exceeded 1, reaching 1.162. This indicates that MLLMs fail to effectively and robustly understand the fine-grained spatial structure of road markings in images. Furthermore, the results of the BEV road network correction task indicate that MLLMs struggle to accurately provide coordinates for corrected intersections and road segments. The best model exhibits an RMSE@20% as high as 0.132 for junction points and an FD@50% as high as 0.232 for road segment polylines, both falling within the same order of magnitude as the upper threshold set. These results suggest that current MLLMs are unable to reason based on fine-grained spatial elements and complete tasks requiring global information and also fully demonstrate that RoadBench is a highly challenging benchmark.

MLLMs struggle to outperform simple rule-based methods that do not rely on any image inputs. Comparing against baselines based on simple rules is a better way to understand the above results than simply interpreting the absolute values of the outcome metrics. Observing the results of the two lane designation recognition tasks reveals that most MLLMs can not outperform the baseline designed based on traffic domain common sense. In other scenarios except the FPV road type classification task, some MLLMs also fail to outperform the baseline based on random choice. By comparing results against simple rules or random selections, we find that MLLMs still have significant room for improvement in fine-grained spatial understanding and the application of domain-specific common sense.

MLLMs struggle to provide precise coordinate numbers. The BEV road network correction task requires MLLMs to return coordinates for points and polylines based on their understanding and reasoning of the input. By examining

the metrics shown in Table 7-3 at the small threshold (10%), we observe that the results are almost entirely contributed by the distance upper threshold. This indicates that the points or lines generated by MLLMs shift significantly from the ground truth, reflecting their limitations in spatial understanding and reasoning, or in the accuracy of structured numerical outputs.

MLLMs are better at correctly understanding the fine-grained spatial elements contained within FPV images. Comparing the results of the same task under FPV and BEV viewpoints, we observe that MLLMs demonstrate significantly superior performance on FPV images compared to BEV images, both in terms of absolute metric values and relative gaps compared to the baseline. This indicates that larger spatial elements such as road markings and signs in the FPV viewpoint can be better understood by MLLMs. Conversely, it also reflects the limitations of MLLMs in understanding more granular spatial elements in BEV images.

Cross-view reasoning remains challenging for current MLLMs. Across the Cross-View Lane Counting and Cross-View Lane Designation tasks, most models exhibit noticeably lower performance compared with single-view tasks. Even strong models such as GPT-5 and Gemini-2.5 show a clear performance drop, indicating that aligning spatial structures between BEV and FPV views remains difficult. Although some models achieve relatively better results, the overall performance gap suggests that robust cross-view spatial understanding is still an open challenge for current multimodal large language models.

Fine-tuning substantially improves the performance of small Qwen3-VL models. After task-specific tuning, Qwen3-VL-2B, Qwen3-VL-4B, and Qwen3-VL-8B show consistent gains across most RoadBench tasks, with noticeable improvements in lane counting, road type classification, and cross-view reasoning. Among them, Qwen3-VL-8B (tuned) achieves the best overall performance within this group and becomes competitive with several larger open-source models. Nevertheless, a performance gap remains compared with leading closed-source models such as GPT-5 and Gemini-2.5, indicating that model scale and proprietary training pipelines still provide notable advantages.

The number of parameters is not a universal solution in RoadBench. Although the number of parameters is largely positively correlated with the capabilities of MLLMs, larger models do not necessarily perform better. For example, GPT-5-Mini outperforms GPT-5 on most tasks. Qwen2.5-VL-32B also outperformed Qwen2.5-VL-72B in the FPV lane recognition task. This may be related to the fusion method or training process of the visual and textual modalities within MLLMs.

Closed-source models hold certain technical advantages on RoadBench. From an overall ranking perspective, in terms of fine-grained spatial understanding capabilities within urban scenarios, closed-source models represented by the GPT-5 series and Gemini-2.5 series have achieved certain technical advantages over open-source models. Among open-source models, only GLM-4.5V ranks highly in comprehensive evaluations.

4.3 Further Analysis

The impact of reference line prompting methods in the BEV tasks. Since reference lines serve as the primary source for MLLMs to identify regions of interest in BEV tasks, how their positions are prompted to MLLMs may directly impact the task performance of MLLMs. We selected the closed-source model GPT-5-Mini and the open-source model GLM-4.5V, which demonstrated better performance in the BEV lane counting task and the BEV lane designation recognition task for further testing. We design different reference line prompt methods, including text-only prompts, visual-only prompts, and both text and visual prompts. Visual prompts are categorized into two approaches: using start and end point colors to indicate direction, and employing arrows on the line segments to denote direction. Both text and visual prompts with arrows are the default settings in the BEV tasks.

Based on the experimental results presented in Table 8-10 and Table 8-11, we observe significant differences in preference for prompt formats among various MLLMs, potentially attributable to variations in the data distribution used during training. GLM-4.5V shows a stronger tendency to learn from both textual prompts and image prompts with arrows, with image prompts playing a dominant role. When deprived of image prompts, GLM-4.5V experiences a noticeable performance decline, whereas removing textual prompts results in only minor fluctuations. GPT-5-Mini shows a stronger preference for prompts that use color to distinguish directions, while different prompting strategies lead to only marginal performance variations in the BEV lane designation recognition task. These phenomena may indicate that GPT-5-Mini is better equipped to integrate textual and visual information to collaboratively process the entire information flow.

The impact of scene environment conditions in the FPV tasks. FPV images captured from in-vehicle cameras encompass varying external environmental conditions. Among these, the most direct factors affecting MLLM understanding and reasoning are adverse lighting conditions and obscured road markings. Regarding obscured road markings, only images where lane information could be determined through alternative means were retained during manual dataset proofreading. After additional annotation, 175 test cases involving adverse lighting conditions were identified, alongside 46 test cases featuring obscured road markings. To analyze the impact of different scenario environments on MLLMs, we selected GLM-4.5V, GPT-5, and Gemini-2.5-Pro which performed well in the main results as case studies.

The experimental results in the FPV lane counting task and the FPV lane designation recognition task are presented in Table 8-12, Table 8-13 and Table 8-14. Based on the results, we have two primary findings. First, adverse lighting conditions do indeed degrade the performance of MLLMs, but this impact is significantly alleviated in models with strong image understanding and reasoning capabilities. Second, shifting the basis for task completion from obscured road markings to other elements like signage substantially improved the performance of both GPT-5 and Gemini-2.5-Pro. This phenomenon indicates that MLLMs

exhibit significantly weaker capabilities in understanding and reasoning fine-grained spatial elements compared to other capabilities.

5 Conclusion

In this paper, we propose a benchmark named RoadBench with eight benchmark tasks and 3,040 test cases for comprehensively evaluating MLLMs' understanding and reasoning of fine-grained spatial elements under urban scenarios based on both BEV and FPV images. Based on this benchmark, we evaluated 20 mainstream MLLMs. The results and further analysis indicate that existing MLLMs lack proper fine-grained spatial understanding and reasoning capabilities under urban scenarios. On certain tasks and metrics, they even fail to outperform baselines based on random selection or simple rules. These findings indicate the significance of RoadBench while also highlighting the need to enhance MLLM's capabilities in fine-grained spatial understanding and reasoning. Based on this fact, RoadBench is promising to become the foundational dataset and evaluation framework for advancing research and applications that enhance the fine-grained spatial understanding and reasoning capabilities of MLLM or MLLM-based agents.

Acknowledgements

This work was supported in part by the National Key Research and Development Program of China under Grant No. 2024YFC3307603, in part by the National Natural Science Foundation of China under Grant No. 62441229. This work was also sponsored by Tsinghua-Toyota Joint Research Institute Inter-disciplinary Program and the Zhongguancun Academy (Grant No. XTS0074).

Data Ethics and Reproducibility. RoadBench is intended solely for academic research and benchmark evaluation. The FPV street-view images are obtained from AMAP with authorization for research use and public release, and all released images have undergone privacy screening, including the removal or blurring of personally identifiable information such as recognizable faces and license plates. We also randomize all sample IDs into UUIDs and convert geographic coordinates into image-level pixel coordinates. The BEV satellite images are based on Google satellite imagery; to support reproducibility while respecting commercial imagery constraints, we provide region-image mappings, geographic coordinates, image filenames, acquisition instructions, construction scripts, prompts, evaluation code, and raw model outputs, enabling users to obtain or reconstruct the images without their own Google API key or paid Google Maps Platform account. Since online map services may update imagery over time, we document the data provenance and provide consistency checks for diagnosing noticeable deviations from the benchmark reference version. RoadBench focuses on five Chinese cities with relatively consistent traffic conventions, and should not be directly used for operational, enforcement, or safety-critical decisions without further ethical review, fairness assessment, and domain-specific validation.

References

1. Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al.: Gpt-4 technical report. arXiv preprint arXiv:2303.08774 (2023) 2, 4
2. Azzolini, A., Bai, J., Brandon, H., Cao, J., Chattopadhyay, P., Chen, H., Chu, J., Cui, Y., Diamond, J., Ding, Y., et al.: Cosmos-reason1: From physical common sense to embodied reasoning. arXiv preprint arXiv:2503.15558 (2025) 4
3. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., et al.: Qwen3-vl technical report. arXiv preprint arXiv:2511.21631 (2025) 10
4. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., et al.: Qwen2. 5-vl technical report. arXiv preprint arXiv:2502.13923 (2025) 2, 4, 10
5. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners. Advances in neural information processing systems **33**, 1877–1901 (2020) 4
6. Chang, X., Xue, M., Liu, X., Pan, Z., Wei, X.: Driving by the rules: A benchmark for integrating traffic sign regulations into vectorized hd map. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 6823–6833 (2025) 5
7. Chen, B., Xu, Z., Kirmani, S., Ichter, B., Sadigh, D., Guibas, L., Xia, F.: Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 14455–14465 (June 2024) 4
8. Chen, J., Wang, H., Li, J., Liu, Y., Dong, Z., Yang, B.: Spatialllm: From multi-modality data to urban spatial intelligence. arXiv preprint arXiv:2505.12703 (2025) 4
9. Cheng, A.C., Yin, H., Fu, Y., Guo, Q., Yang, R., Kautz, J., Wang, X., Liu, S.: Spatialrgpt: Grounded spatial reasoning in vision-language models. Advances in Neural Information Processing Systems **37**, 135062–135093 (2024) 4
10. Comanici, G., Bieber, E., Schaekermann, M., Pasupat, I., Sachdeva, N., Dhillon, I., Blistein, M., Ram, O., Zhang, D., Rosen, E., et al.: Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. arXiv preprint arXiv:2507.06261 (2025) 10
11. Cui, C., Ma, Y., Cao, X., Ye, W., Zhou, Y., Liang, K., Chen, J., Lu, J., Yang, Z., Liao, K.D., et al.: A survey on multimodal large language models for autonomous driving. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. pp. 958–979 (2024) 2, 4
12. Daxberger, E., Wenzel, N., Griffiths, D., Gang, H., Lazarow, J., Kohavi, G., Kang, K., Eichner, M., Yang, Y., Dehghan, A., et al.: Mm-spatial: Exploring 3d spatial understanding in multimodal llms. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7395–7408 (2025) 5
13. Driess, D., Xia, F., Sajjadi, M.S., Lynch, C., Chowdhery, A., Ichter, B., Wahid, A., Tompson, J., Vuong, Q., Yu, T., et al.: Palm-e: An embodied multimodal language model. In: International Conference on Machine Learning. pp. 8469–8488. PMLR (2023) 2
14. Eiter, T., Mannila, H., et al.: Computing discrete fréchet distance (1994) 7
15. Feng, J., Liu, T., Du, Y., Guo, S., Lin, Y., Li, Y.: 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. pp. 591–602 (2025) 2, 4

16. Feng, J., Wang, S., Liu, T., Xi, Y., Li, Y.: Urbanllava: A multi-modal large language model for urban intelligence with spatial reasoning and understanding. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6209–6219 (2025) 2, 4
17. Feng, J., Zhang, J., Liu, T., Zhang, X., Ouyang, T., Yan, J., Du, Y., Guo, S., Li, Y.: Citybench: Evaluating the capabilities of large language models for urban tasks. In: Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2. pp. 5413–5424 (2025) 2, 5
18. Fu, C., Chen, P., Shen, Y., Qin, Y., Zhang, M., Lin, X., Yang, J., Zheng, X., Li, K., Sun, X., Wu, Y., Ji, R.: Mme: A comprehensive evaluation benchmark for multimodal large language models (2024), https://arxiv.org/abs/2306.13394 4
19. Guo, Q., De Mello, S., Yin, H., Byeon, W., Cheung, K.C., Yu, Y., Luo, P., Liu, S.: Regiongpt: Towards region understanding vision language model. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13796–13806 (2024) 4
20. Hu, H., Wang, P., Bi, H., Tong, B., Wang, Z., Diao, W., Chang, H., Feng, Y., Zhang, Z., Wang, Y., et al.: Rs-vheat: Heat conduction guided efficient remote sensing foundation model. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9876–9887 (2025) 4
21. Li, B., Ge, Y., Ge, Y., Wang, G., Wang, R., Zhang, R., Shan, Y.: Seed-bench: Benchmarking multimodal large language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13299–13308 (2024) 4
22. Li, J., Li, D., Xiong, C., Hoi, S.: Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In: International conference on machine learning. pp. 12888–12900. PMLR (2022) 4
23. Li, X., Zhang, M., Geng, Y., Geng, H., Long, Y., Shen, Y., Zhang, R., Liu, J., Dong, H.: Manipllm: Embodied multimodal large language model for object-centric robotic manipulation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18061–18070 (2024) 2
24. Liu, Y., Duan, H., Zhang, Y., Li, B., Zhang, S., Zhao, W., Yuan, Y., Wang, J., He, C., Liu, Z., et al.: Mmbench: Is your multi-modal model an all-around player? In: European conference on computer vision. pp. 216–233. Springer (2024) 4
25. Lu, P., Bansal, H., Xia, T., Liu, J., Li, C., Hajishirzi, H., Cheng, H., Chang, K.W., Galley, M., Gao, J.: Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. arXiv preprint arXiv:2310.02255 (2023) 4
26. OpenDataLab: Labelu (2025), https://github.com/opendatalab/labelU-kit, accessed: 2025-09-01 10
27. Ou, R., Hu, Y., Zhang, F., Chen, J., Liu, Y.: Geopix: A multimodal large language model for pixel-level image understanding in remote sensing. IEEE Geoscience and Remote Sensing Magazine (2025) 2, 4
28. Park, S.Y., Cui, C., Ma, Y., Moradipari, A., Gupta, R., Han, K., Wang, Z.: Nuplanqa: A large-scale dataset and benchmark for multi-view driving scene understanding in multi-modal large language models. arXiv preprint arXiv:2503.12772 (2025) 5
29. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021) 4
30. Ramakrishnan, S.K., Wijmans, E., Kraehenbuehl, P., Koltun, V.: Does spatial cognition emerge in frontier models? arXiv preprint arXiv:2410.06468 (2024) 4

31. Roberts, J., Lüddecke, T., Sheikh, R., Han, K., Albanie, S.: Charting new territories: Exploring the geographic and geospatial capabilities of multimodal llms. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 554–563 (2024) 2
32. Team, G., Kamath, A., Ferret, J., Pathak, S., Vieillard, N., Merhej, R., Perrin, S., Matejovicova, T., Ramé, A., Rivière, M., et al.: Gemma 3 technical report. arXiv preprint arXiv:2503.19786 (2025) 2, 4, 10
33. Team, V., Hong, W., Yu, W., Gu, X., Wang, G., Gan, G., Tang, H., Cheng, J., Qi, J., Ji, J., Pan, L., Duan, S., Wang, W., Wang, Y., Cheng, Y., He, Z., Su, Z., Yang, Z., Pan, Z., Zeng, A., Wang, B., Chen, B., Shi, B., Pang, C., Zhang, C., Yin, D., Yang, F., Chen, G., Xu, J., Zhu, J., Chen, J., Chen, J., Chen, J., Lin, J., Wang, J., Chen, J., Lei, L., Gong, L., Pan, L., Liu, M., Xu, M., Zhang, M., Zheng, Q., Yang, S., Zhong, S., Huang, S., Zhao, S., Xue, S., Tu, S., Meng, S., Zhang, T., Luo, T., Hao, T., Tong, T., Li, W., Jia, W., Liu, X., Zhang, X., Lyu, X., Fan, X., Huang, X., Wang, Y., Xue, Y., Wang, Y., Wang, Y., An, Y., Du, Y., Shi, Y., Huang, Y., Niu, Y., Wang, Y., Yue, Y., Li, Y., Zhang, Y., Wang, Y., Wang, Y., Zhang, Y., Xue, Z., Hou, Z., Du, Z., Wang, Z., Zhang, P., Liu, D., Xu, B., Li, J., Huang, M., Dong, Y., Tang, J.: Glm-4.5v and glm-4.1v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning (2025), https://arxiv.org/abs/2507.01006 10
34. Tian, X., Gu, J., Li, B., Liu, Y., Wang, Y., Zhao, Z., Zhan, K., Jia, P., Lang, X., Zhao, H.: Drivevlm: The convergence of autonomous driving and large vision-language models. In: Conference on Robot Learning. pp. 4698–4726. PMLR (2025) 2
35. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al.: Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971 (2023) 4
36. Wang, J., Ming, Y., Shi, Z., Vineet, V., Wang, X., Li, S., Joshi, N.: Is a picture worth a thousand words? delving into spatial reasoning for vision language models. Advances in Neural Information Processing Systems **37**, 75392–75421 (2024) 4
37. Wang, X., Ma, W., Zhang, T., de Melo, C.M., Chen, J., Yuille, A.: Spatial457: A diagnostic benchmark for 6d spatial reasoning of large multimodal models. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 24669–24679 (2025) 4
38. Wei, H., Yin, Y., Li, Y., Wang, J., Zhao, L., Sun, J., Ge, Z., Zhang, X., Jiang, D.: Slow perception: Let's perceive geometric figures step-by-step. arXiv preprint arXiv:2412.20631 (2024) 4
39. Xiao, H., Zhou, F., Liu, X., Liu, T., Li, Z., Liu, X., Huang, X.: A comprehensive survey of large language models and multimodal large language models in medicine. Information Fusion p. 102888 (2024) 4
40. Xie, S., Kong, L., Dong, Y., Sima, C., Zhang, W., Chen, Q.A., Liu, Z., Pan, L.: Are vlms ready for autonomous driving? an empirical study from the reliability, data, and metric perspectives. arXiv preprint arXiv:2501.04003 (2025) 2, 5
41. Xuan, W., Wang, J., Qi, H., Chen, Z., Zheng, Z., Zhong, Y., Xia, J., Yokoya, N.: Dynamicvl: Benchmarking multimodal large language models for dynamic city understanding. arXiv preprint arXiv:2505.21076 (2025) 4
42. Yang, J., Yang, S., Gupta, A.W., Han, R., Fei-Fei, L., Xie, S.: Thinking in space: How multimodal large language models see, remember, and recall spaces. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 10632–10643 (2025) 4

43. Yin, S., Fu, C., Zhao, S., Li, K., Sun, X., Xu, T., Chen, E.: A survey on multimodal large language models. National Science Review **11**(12), nwae403 (2024) 2, 4
44. Zhang, R., Zhang, W., Tan, X., Yang, S., Wan, X., Luo, X., Li, G.: Vldrive: Vision-augmented lightweight mllms for efficient language-grounded autonomous driving. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5923–5933 (2025) 4
45. Zhang, S., Chen, A., Sun, Y., Gu, J., Zheng, Y.Y., Koniusz, P., Zou, K., Hengel, A.v.d., Xue, Y.: Open eyes, then reason: Fine-grained visual mathematical understanding in mllms. arXiv preprint arXiv:2501.06430 (2025) 2, 4
46. Zhang, W., Cai, M., Zhang, T., Zhuang, Y., Li, J., Mao, X.: Earthmarker: A visual prompting multi-modal large language model for remote sensing. IEEE Transactions on Geoscience and Remote Sensing (2024) 2
47. Zhang, X., Ouyang, T., Shang, Y., Liao, Q., Li, Y.: UrbanMLLM: Joint learning of cross-view imagery for urban understanding. In: Forty-third International Conference on Machine Learning (2026), https://openreview.net/forum?id=Av1OFtirO4 2, 4
48. Zhang, Y., Nie, Y.: Interndrive: A multimodal large language model for autonomous driving scenario understanding. In: Proceedings of the 2024 4th International Conference on Artificial Intelligence, Automation and High Performance Computing. pp. 294–305 (2024) 2
49. Zhao, B., Wang, Z., Fang, J., Gao, C., Man, F., Cui, J., Wang, X., Chen, X., Li, Y., Zhu, W.: Embodied-r: Collaborative framework for activating embodied spatial reasoning in foundation models via reinforcement learning. In: Proceedings of the 33rd ACM International Conference on Multimedia. pp. 11071–11080 (2025) 4
50. Zhao, Y., Xu, K., Zhu, Z., Hu, Y., Zheng, Z., Chen, Y., Ji, Y., Gao, C., Li, Y., Huang, J.: Cityeqa: A hierarchical llm agent on embodied question answering benchmark in city space. EMNLP (2025) 5
51. Zhou, B., Yang, H., Chen, D., Ye, J., Bai, T., Yu, J., Zhang, S., Lin, D., He, C., Li, W.: Urbench: A comprehensive benchmark for evaluating large multimodal models in multi-view urban scenarios. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 10707–10715 (2025) 2, 5