

# Fine-Grained Evaluation of Large Vision-Language Models in Autonomous Driving

Yue Li<sup>1\*</sup> Meng Tian<sup>2</sup> Zhenyu Lin<sup>2</sup> Jiangtong Zhu<sup>2</sup> Dechang Zhu<sup>2</sup>  
 Haiqiang Liu<sup>2</sup> Yueyi Zhang<sup>1</sup> Zhiwei Xiong<sup>1,✉</sup> Xinhai Zhao<sup>2,✉</sup>  
<sup>1</sup> University of Science and Technology of China <sup>2</sup> Huawei Noah's Ark Lab  
 yueli65@mail.ustc.edu.cn zwxiong@ustc.edu.cn zhaoxinhail@huawei.com

## Abstract

Existing benchmarks for Vision-Language Model (VLM) in autonomous driving (AD) primarily assess interpretability through open-form visual question answering (QA) within coarse-grained tasks, which remain insufficient to assess capabilities in complex driving scenarios. To this end, we introduce **VLADBench**, a challenging and fine-grained benchmark featuring close-form QAs that progress from static foundational knowledge and elements to advanced reasoning for dynamic on-road situations. The elaborate **VLADBench** spans 5 key domains: Traffic Knowledge Understanding, General Element Recognition, Traffic Graph Generation, Target Attribute Comprehension, and Ego Decision-Making and Planning. These domains are further broken down into 11 secondary aspects and 29 tertiary tasks for a granular evaluation. A thorough assessment of general and domain-specific (DS) VLMs on this benchmark reveals both their strengths and critical limitations in AD contexts. To further exploit the cognitive and reasoning interactions among the 5 domains for AD understanding, we start from a small-scale VLM and train the DS models on individual domain datasets (collected from 1.4M DS QAs across public sources). The experimental results demonstrate that the proposed benchmark provides a crucial step toward a more comprehensive assessment of VLMs in AD, paving the way for the development of more cognitively sophisticated and reasoning-capable AD systems. The benchmark is available at <https://github.com/Depth2World/VLADBench>.

## 1. Introduction

Large Vision-Language Models (VLMs) are rapidly transforming numerous fields, demonstrating their potential to revolutionize how we interact with information and technol-

\*The work was done during Yue Li's internship at Huawei Noah's Ark Lab.

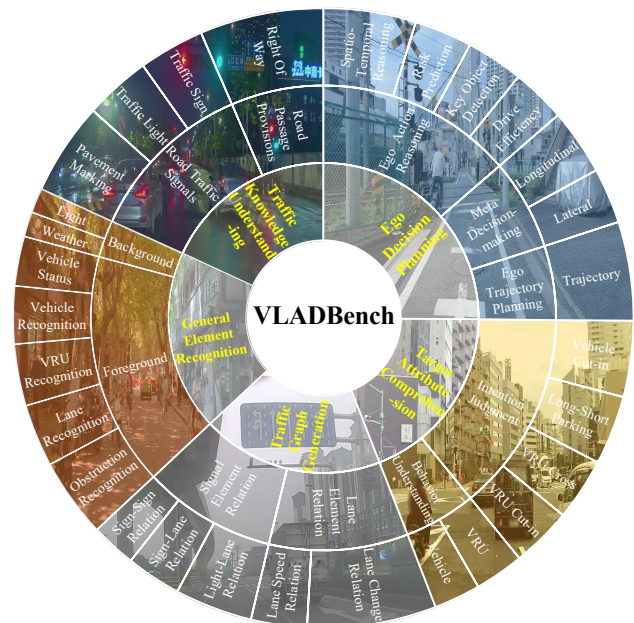


Figure 1. A sunburst chart of **VLADBench** categories. The proposed benchmark spans 5 key domains, 11 secondary aspects and 29 tertiary tasks, including about 2,000 static scenes and 3,000 dynamic scenarios, comprising 12,000 close-form questions.

ogy. Their ability to seamlessly integrate visual and textual data unlocks new possibilities across diverse applications, including visual content generation [2, 34, 39], medical image analysis [10, 48, 74], robotic control [26, 81], and autonomous driving (AD) [14, 15, 58, 61, 71].

Recent VLM-based AD algorithms address the limitations of end-to-end AD approaches, including interpretability and long-tail problem, which refers to the limited generalization to new scenarios, unexpected events, and diverse traffic patterns [9, 60]. State-of-the-art models [46, 47, 49, 58, 61, 71] demonstrate promising results in scene perception, description, and decision-making with analysis in open-form visual question answer (VQA) task. Most existing datasets focus primarily on high-level task categories

such as perception, prediction, and planning in AD.

Despite recent advances, the effective transfer of foundational VLMs to AD-specific models remains underexplored, in part due to the insufficient validation protocols within the context of AD. A comprehensive evaluation framework is essential to guide the model transfer by identifying not only overall performance but also the specific strengths and limitations of each model, going beyond high scores on open-form VQA tasks. Current benchmarks designed for VLM-based AD face several notable limitations: 1) **Coarse-grained Categories**: The underlying datasets of the VLM-based models are often simplistic, typically categorizing tasks into perception, prediction, and planning with reasoning, which are incomplete for evaluating the nuanced cognitive and reasoning abilities required for safe and reliable AD. A holistic evaluation framework remains necessary to fully assess these critical competencies. 2) **Limited Analysis of Dynamic Elements**: Both static and dynamic scenes are crucial for evaluating AD systems, a robust analysis of dynamic elements is particularly important for validating the temporal reasoning capabilities, especially in understanding traffic participant intentions within the scene and executing the nuanced spatio-temporal reasoning required for safe navigation. 3) **Lack of Data Diversity**: Existing AD benchmark datasets are overly homogeneous, limiting their utility for testing generalization across diverse real-world scenarios. The narrow results restrict the evaluation of zero-shot generalization and the performance on challenging corner cases. A more diverse dataset is required to thoroughly assess the robustness and adaptability of VLMs in real-world settings.

To overcome these limitations, we introduce a novel benchmark, **VLADBench**, specifically designed to rigorously evaluate the capabilities of VLMs in AD. **VLADBench** addresses the shortcomings of existing benchmarks by employing a hierarchical structure that reflects the complex skill set required for reliable driving, progressing from fundamental scene and traffic elements comprehension to advanced reasoning and decision-making. With 2,000 static scenes and 3,000 dynamic scenarios, **VLADBench** spans 5 primary domains: Traffic Knowledge Understanding (TKU), General Element Recognition (GER), Traffic Graph Generation (TGG), Target Attribute Comprehension (TAC), and Ego Decision-making and Planning (EDP). For a more detailed assessment, 11 secondary aspects and 29 tertiary tasks are defined, resulting in a total of 12,000 questions. The benchmark is built from existing publicly available datasets, meticulously curated through a manual selection across 12 sources, aimed at challenging VLM capabilities in diverse challenging driving situations. To further investigate the intersections among the 5 key domains and their contributions for motion planning, we curate a domain-specific (DS) training dataset compris-

ing approximately 1.4M QA pairs collected from public resources. These QA pairs are categorized into the five domains using GPT-4. We then train models on each individual DS datasets and evaluate their performance on **VLADBench** to assess their capabilities across different domains.

A thorough evaluation on **VLADBench** of the prominent VLMs, encompassing both open-source (ranging from 4B to 76B), close-source and DS models, reveals the following key findings:

- Among current VLMs, only Qwen2.5-VL-72B[4] surpasses a score of 60 on **VLADBench**, while others (including GPT-4o) struggle to reach this threshold, indicating substantial room for improvement.
- Significant challenges persist especially in areas: traffic signals and graph generation, intention judgment, and meta decision-making, which are essential capabilities for achieving reliable AD.
- Biased DS training data improve performance in certain specialized areas of AD but often compromise generalization ability in tasks that require broader and more general knowledge.
- The DS data from the five key domains is interconnected, providing mutual benefits across domains and demonstrating a clear synergy effect.
- Elevating the vision encoder may be more impactful than simply scaling up the language model for AD context.

## 2. Related Work

### 2.1. Large Vision-Language Models

Recent advancements in Large Language Models (LLMs) like the GPT series [1] and LLaMA [63] have revolutionized natural language processing. This progress has spurred the development of large vision-language models, aiming to extend LLM capabilities to encompass visual understanding and reasoning. Models such as LLaVA [39, 40], MiniGPT-4 [80], InstructBLIP [13], Cambrian [62], ShareGPT4V [8] integrate visual information, enabling tasks like image captioning and visual question answering. These VLMs typically align visual and linguistic features using cross-attention mechanisms or MLP projections, trained on extensive image-text datasets. Early VLMs focused on static images, but recent efforts have extended their capabilities to video understanding, such as BLIP2 [32], InternLM-XComposer2.5 [77], InternVL2 [12], VILA [38], Qwen2-VL [66], etc., incorporating temporal dynamics into the language feature space for sequence comprehension. VLMs have demonstrated promising capabilities across diverse domains, including content creation, medical image analysis, robotics and autonomous driving.

Table 1. Comparison between the existing benchmarks and our proposed benchmark. V. and Cate. represent video and category.

Benchmark	Source	V.	QAs	Cate.
CODA-LM [35]	CODA [33]	×	1.5K	3
LingoQA [47]	Self-collected	✓	1K	4
IDKB [42]	Internet	✓	20K	4
nuScenes-QA [54]	nuScenes [6]	✓	83K	5
DriveLM [58]	nuScenes [6]	✓	15K	4
DriveBench [70]	nuScenes [6]	✓	21K	5
MME-Realworld [79]	[6, 33, 56]	×	5K	15
NuInstruct [15]	nuScenes [6]	✓	16K	17
	[6, 44, 45, 68]			
<b>VLADBench</b>	[18, 21, 28, 33]	✓	12K	29
	[19, 29, 47, 55]			

## 2.2. VLM-based Autonomous Driving

End-to-end AD [22, 24] represents a shift from traditional modular pipelines to a singular framework, which learns relevant features directly from raw sensor data and discovers effective representations with all modules training together. While models trained on specific datasets encounter the reliance on ego status [36, 75] and long-tail dilemma, i.e., fail to generalize on new scenarios, unexpected events, or traffic patterns [9, 60]. Besides, these approaches typically lack interpretability, making it difficult to explain their actions and hindering trust and regulatory approval.

To address these problems, several recent works explore the potential of VLMs for AD. LingoQA [47] and Dolphins [43], for example, employ VQA to bridge the gap between data-driven driving and user trust. Besides, decision-making and planning are also being integrated into VLMs, as seen in DriveVLM [61], DriveLM [58], Reason2drive [49], BEV-InMLMM [15], OmniDrive [67], where the training data is always divided into perception, prediction, and planning components. These models often produce outputs via a chain-of-thought (CoT) process, encompassing scene descriptions, action analysis, hierarchical planning, etc. Approaches such as DriveGPT4 [71], VLP [50], AsyncDriver [11] and LMDrive [57] attempt to directly map visual and linguistic inputs to planning or low-level control signals. The end-to-end AD systems based on VLMs offer strong interpretability, trustworthiness, and the ability to understand complex scenes.

## 2.3. Benchmark and Metrics

Established evaluation benchmarks like MME [16], VideoMME [17], MMBench [41] and Seed-Bench [30], while valuable for foundation models, are not ideally suited for evaluating AD models, because that these benchmarks primarily comprise natural images, lacking the specific characteristics of driving scenarios, such as traffic elements and the dynamic interactions of participants. Recent works have introduced specified AD datasets with extensive VQA

Table 2. Prompt setting of **VLADBench**. \* denotes optional.

<b>Most Question</b>
[Image / Video] [Visual Prompt]* [Question] [Tips]*. Select one as the answer from the list below: [Choice A, B, C, D, E]. No explanation is needed. The best answer is:
<b>Other Question: Detection, Traffic Graph, Trajectory</b>
[Image / Video] [Visual Prompt]* [Question] [Tips]*. [Output Format]. No explanation is needed. The answer is:

pairs. Nuscenes-QA [54], CODA-LM [35], VLAAD [53] and LingoQA [47] start from scene description and analysis, general perception, action reasoning and driving suggestions. DriveLM [58], NuInstruct [15], Reason2Drive [49] divide the data into perception, prediction, and planning with reasoning. DriveLM [58] also includes behavior understanding, and NuInstruct [15] includes risk estimation. IDKB [42] mined plenty of questions about 4 traffic knowledge domains from various handbooks. DriveBench [70] further introduce corruption data for robustness validation. With the rapid advancement, a coarse categorization is insufficient to support a complete analysis of AD models.

For evaluation, language-based metrics like BLEU [51], ROUGE [37], METEOR [5] and CIDEr [64], commonly used to evaluate question-answering models, however, demonstrate poor correlation with human judgment. This is problematic because semantically distinct sentences with opposite meanings also can receive similar scores, posing unacceptable risks in safety-critical AD applications. While recent metrics leveraging ChatGPT ratings [35, 71], they exhibit positional and stylistic biases and produce inconsistent scores across iterations. In this paper, we revisit the simple yet effective metric: Accuracy. Through the close-form instruction annotations, we try to achieve a precise evaluation in terms of the zero tolerance for error.

## 3. Benchmark

### 3.1. Data Source and Annotation

**Data Source.** A comprehensive and diverse benchmark is able to reduce testing bias, which helps probe a thorough evaluation of zero-shot generalization capabilities and better expose the weakness of VLMs in various AD scenarios. As shown in Tab. 1, existing benchmarks often suffer from a lack of diversity. In contrast, our proposed benchmark **VLADBench**, covering 5 domains, 11 aspects and 29 tasks, constructed from the 12 publicly available datasets: GTSDDB [21], JAAD [29], PIE [55], HAD [28], nuScenes [6], SODA [19], ONCE [45], Argoverse2 [68], CODA and CODA2022 [33], DRAMA [44], RS10K [18], and LingoQA [47]. The instance counts for the five domains are: TKU (2,369), GER (2,812), TGG (3,090), TAC

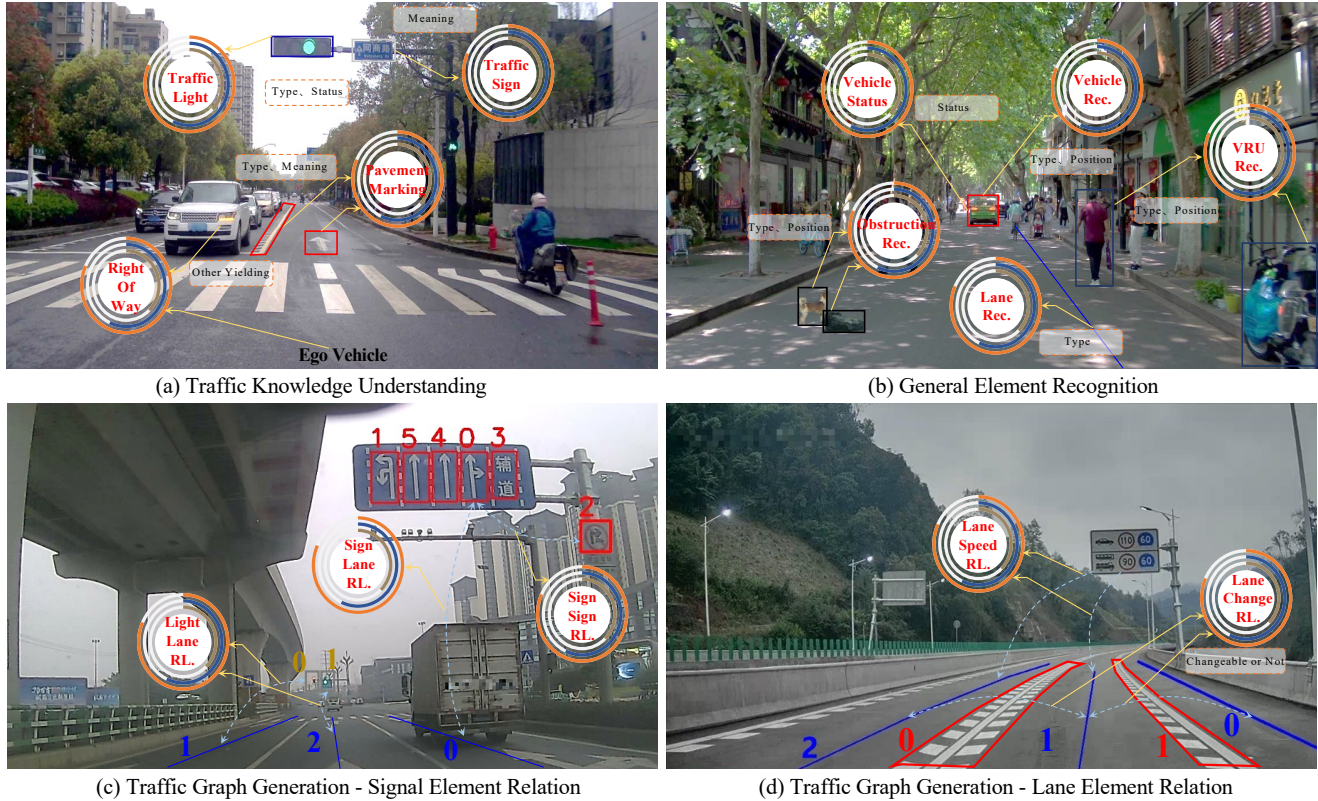


Figure 2. Real-world examples of the tasks in (a) Traffic Knowledge Understanding, (b) General Element Recognition, and (c, d) Traffic Graph Generation domains. ‘Rec.’ and ‘RL’ denote recognition and relation.

(1,303), and EDP (2,418). Detailed numbers of each task are included in the supplement.

**Annotation.** Based on the designed domains and tasks, we meticulously hand-select 2,000 static and 3,000 dynamic scenes for a diverse range of challenging driving situations. During the selection process, we control the visual prominence of objects and scenes to avoid immediate recognition. Existing datasets predominantly feature object- or caption-level annotations, lacking detailed and task-specific annotations. Consequently, we engage 5 human annotators for fine-grained annotation and implement a quality double-check with 2 professional researchers. Each instance takes about 5 minutes to annotate.

### 3.2. Instruction and Criterion

**Instruction.** For most of the questions in the proposed benchmark, we first construct each QA pair and then we collect all the answers in each task as a database. After that, we select the correct answer and randomly select the incorrect answers to form a choice list for each question. The choices in the list are semantically or structurally similar, increasing the ambiguity and difficulty of the question. The instruction format is listed in Tab. 2 and the length of the list ranges from 4 to 10. For the other types of questions, i.e., visual detection in GER, TGG and trajectory planning, we specify the output format for each question to guide the in-

struction following. Some questions in GER and TGG are constructed with visual prompts and descriptive tips. The visual prompts include the bounding boxes on the image or the coordinates of these boxes in the instructions, which are employed for regional perception representations. Besides, some questions in TGG are also constructed with tips comprising the perceptual descriptions within the scene, aiming at assisting the challenging task by providing accurate perceptual prior.

**Criterion.** For the evaluation of each task, the core metrics are accuracy and instruction compliance rate. Besides, there are IOU for detection in the recognition tasks, judgment accuracy for the intention judgment tasks, and L2 distance and collision rate for the ego trajectory planning task. The final score for each task is weighted by these metrics. Note that a rule-based filter is employed to align the responses generated by VLMs with the choice list in the instruction, removing symbols and special tokens.

### 3.3. Data Statistics

**Traffic Knowledge Understanding.** This domain comprises two primary aspects: *Road Traffic Signals* (encompassing the pavement marking, traffic sign, and traffic light tasks, with questions pertaining to type, status, meaning, and optical character recognition) and *Road Passage Provisions* (determining the right-of-way between ego vehicle

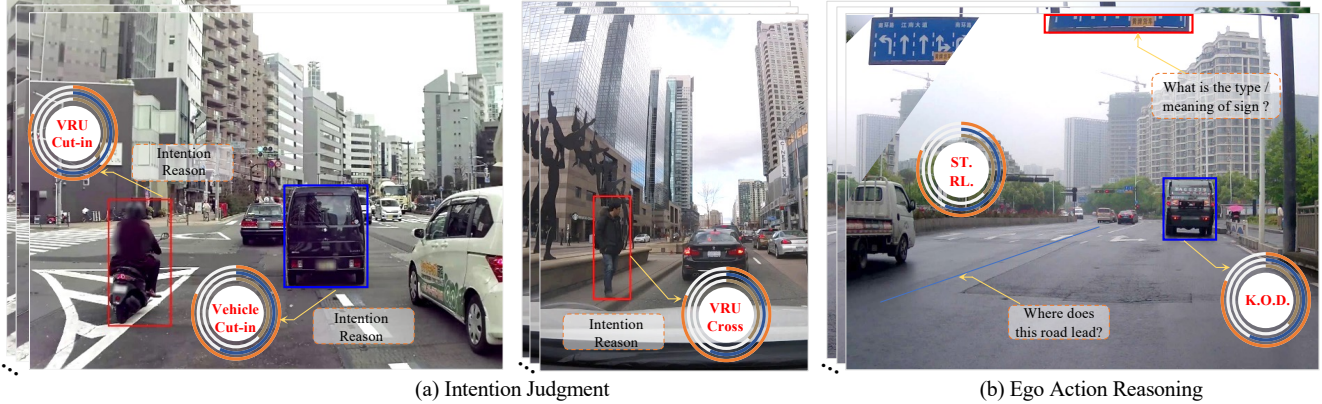


Figure 3. Examples in *intention judgment* and *ego action reasoning* aspects. ST.RL.: spatio-temporal reasoning, K.O.D.: key object detection.

and other traffic participants), as illustrated in Fig. 2 (a).

**General Element Recognition.** This domain consists of *Background* and *Foreground* aspects. *Background* includes light and weather tasks, while *Foreground* focuses on lane recognition, vehicle recognition, vehicle status, vulnerable road user (VRU) recognition, and obstruction recognition tasks. The recognition tasks also involve visual grounding questions, such as instructions with coordinates or object detection. The vehicle status task identifies the external states of the vehicle, such as brake lights, open doors, and trunk. The obstruction (including animals) recognition task, also addresses whether obstacles can be safely driven over, with examples illustrated in Fig. 2 (b).

**Traffic Graph Generation.** The aforementioned assessment allows for the perception of low-level scene elements, which forms the basis for a high-level understanding of the interrelationships between traffic elements, termed traffic graph generation. This domain encompasses both *Signal Element Relation* and *Lane Element Relation* aspects. The former, as illustrated in Fig. 2 (c), refers to the correspondence between the different signal elements in the scene, i.e., light-lane relation, sign-lane relation and sign-sign relation tasks. The latter, as illustrated in Fig. 2 (d), refers to the correspondence between the lane elements in the scene, i.e., lane speed relation and lane change relation tasks.

**Target Attribute Comprehension.** The comprehension of dynamic scenes is paramount in AD. In this domain, we include the temporal analysis by incorporating the prediction of future (unoccurred) events, termed *Intention Judgment* aspect, and the analysis of the past (occurred) events, termed *Behavior Understanding* aspect. As visualized in Fig. 3 (a), *Intention Judgment* is composed of vehicle cut-in, VRU cut-in, VRU cross, and long-short parking tasks, with the questions pertaining to the underlying intention and motivation. *Behavior Understanding* involves analyzing vehicle and VRU behavior by describing the sequence of temporal events. Moreover, we include pedestrian gesture analysis related to right-of-way determination, further

testing human-vehicle interaction capabilities.

**Ego Decision-Making and Planning.** We construct this domain in a reasoning mechanism from *Ego Action Reasoning*, the high-level *Meta Decision-Making* and the final *Ego Trajectory Planning* aspects. *Ego Action Reasoning* contains the fine-grained tasks: key object detection, drive efficiency, risk prediction, and spatio-temporal reasoning, all of which significantly influence subsequent driving strategies. Specifically, in spatio-temporal reasoning, we devise the challenge of inferring the state or meaning of part-occluded traffic signals, lane type, and lane destination at the end of a video sequence. At the moment of the last frame, information from previous frames, such as traffic signs and pavement markings, must be integrated and reasoned upon to answer these questions, as Fig. 3 (b) showcasing the real-world samples. *Meta Decision-Making* focuses on the short-term lateral and longitudinal decisions tasks, which are tactical and on immediate execution. The decisions include but are not limited to straight, changing lane to the left/right, in-lane left/right avoidance, borrowing lane for left/right avoidance, accelerating, stop, maintaining, decelerating, and decelerating to stop. *Ego Trajectory Planning* is formulated as a vision and language task, given the critical perception and prediction results, along with high-level decisions. Besides, the ego status and the historical waypoints (last 2 seconds, given by four points) are included in the instruction. The VLMs then generate a feasible 3-second driving trajectory with 6 waypoints.

## 4. Experiments

### 4.1. Baselines and Settings

**Baselines.** To conduct a comprehensive evaluation on static and dynamic scenes, we compare 20 VLM models including the foundational and domain-specific models, which can be divided into the open-source VLMs: VILA-U (VU) [69], InternLM-XComposer2.5 (IXC2.5) [77], Openflamingo [2], CogVLM2 (CV) [20], LongVILA (LoV) [72],

Table 3. Scores evaluated on **VLADBench** from different VLMs. TKU: Traffic Knowledge Understanding, GER: General Element Recognition, TGG: Traffic Graph Generation, TAC: Target Attribute Comprehension, EDP: Ego Decision-Making and Planning. The gray, yellow and purple cell color denotes the open-source, closed-source, and domain-specific VLMs. The best score for each aspect in red. The detailed results about the 29 tasks are listed in the supplement. Note our baseline is excluded for comparison with existing models.

Domains	Aspects	IXC2.5	CV	LoV	QW	IVL2	MCV	IVL2	QW2	OV	QW2.5	LV	GEM	GPT	Senna	Dols	DriLM	DriMM	DriLM-B	Ours
		8B	8B	8B	7B	4B	8B	8B	7B	7B	7B	7B	1.5pro	4o	7B	9B	4B	7B	4B	4B
TKU	<i>Road Traffic Signals</i>	24.89	47.00	37.46	47.40	48.97	43.91	54.97	54.77	56.89	62.45	57.49	67.56	69.09	10.29	28.39	55.04	57.15	52.56	65.65
	<i>Road Passage Provisions</i>	32.69	59.35	49.45	79.22	80.58	22.72	69.45	71.52	70.81	80.32	74.11	42.98	78.96	15.53	21.10	81.36	42.33	73.85	80.58
GER	<i>Background</i>	22.54	58.79	63.93	63.97	66.83	70.76	70.49	71.07	69.11	69.29	69.11	65.71	68.35	25.36	58.75	64.46	70.31	68.21	71.61
	<i>Foreground</i>	26.20	29.11	38.16	38.13	51.90	44.31	49.34	50.88	53.99	53.64	53.04	51.64	53.82	15.51	29.24	52.80	53.27	51.68	60.47
TGG	<i>Signal Element Relation</i>	14.67	25.99	37.04	23.64	29.09	34.93	37.07	36.93	34.70	32.56	33.83	44.58	49.56	3.37	26.08	26.19	36.01	35.00	47.64
	<i>Lane Element Relation</i>	36.34	54.08	43.65	37.03	45.72	57.73	46.98	48.06	64.60	48.91	66.90	70.43	65.71	52.08	31.13	21.19	23.63	51.19	54.01
TAC	<i>Intention Judgment</i>	67.47	38.98	34.76	68.23	41.56	60.08	46.79	60.62	57.60	53.42	59.52	55.95	47.79	52.79	57.64	47.13	43.27	60.89	52.97
	<i>Behavior Understanding</i>	30.61	33.74	17.99	28.60	44.13	37.54	46.82	41.79	41.68	42.23	43.13	50.61	52.63	12.96	0.11	40.78	40.11	42.91	42.91
EDP	<i>Ego Action Reasoning</i>	45.91	36.89	52.03	64.24	46.47	58.85	61.91	54.93	47.35	58.87	56.50	61.20	65.77	25.70	58.72	55.95	52.99	56.60	69.73
	<i>Meta Decision-Making</i>	55.60	22.98	18.45	23.87	47.62	35.65	40.83	35.24	36.61	35.48	41.19	56.43	48.04	15.00	13.69	37.26	50.00	46.31	57.14
	Overall	32.34	38.16	40.06	44.78	46.21	47.28	50.27	51.07	52.15	52.30	53.63	57.19	58.92	21.34	34.97	45.75	47.01	51.32	59.45

Qwen-VL (QW)[3], MiniCPM-V-2.6 (MCV) [73], InternVL2 (IVL2) [12], Qwen2-VL (QW2) [66], OneVision (OV) [31], LLaVA-Video (LV) [78], Qwen2.5-VL (QW2.5)[4], the closed-source VLMs: Gemini-1.5-pro (GEM) [59], GPT-4o<sup>1</sup>, and the domain-specific VLMs: Dolphins (Dols) [43], Senna [25] (VLM part), DriveLM [58] (DriLM) and DriLM-B (trained on BDD [27]), and DriveMM (DriMM) [23].

**Settings.** For the sequence samples, we adjust the frames extraction to ensure all the frames are fed into the evaluation models. Besides, system prompts are not used, even for models that support them. As mentioned above, each task employs 2 to 3 metrics, which are weighted to compute the final score. Instruction compliance is weighted at 0.2, while accuracy is weighted differently: 0.8 for most tasks, 0.5 for tasks in TGG, and 0.7 for tasks in *intention judgment* aspect. The mean score for each aspect is then computed as the weighted average of its task scores, with weights proportional to the number of tasks in that aspect.

**Domain Data for Training.** To further exploit the interactions among the 5 key domains for AD understanding, we start from a small-scale VLM, IVL2-4B [12], and train the DS models on individual domain datasets. These training datasets, sourced from [7, 18, 28, 33, 42–44, 46, 47, 52, 58, 65, 71], contain a total of 1.4M QAs, covering perspectives from the ego vehicle, including single-view, sequential single-view, and multi-view. The type of each QA is classified using GPT-4. Besides, we also incorporate 1.3M QAs from general data for avoiding general ability loss. The IVL2-4B [12], trained on 2.7M QAs, serves as our baseline in this paper. More details are provided in the supplement.

## 4.2. Experimental Results

First, we assess the existing open-source, closed-source, and DS models. The qualitative results across 10 aspects of

<sup>1</sup><https://openai.com/index/hello-gpt-4o/>

Table 4. Scores evaluated on large-scale VLMs.

Aspects	VU	OV	LV	IVL2	QW2	QW2.5
	40B	72B	72B	76B	72B	72B
<i>Road Traffic Signals</i>	32.32	54.15	59.09	59.94	68.31	70.76
<i>Road Passage Provisions</i>	46.34	76.96	80.06	71.07	81.36	80.58
<i>Background</i>	65.13	70.67	71.43	72.68	70.00	71.96
<i>Foreground</i>	42.37	51.91	52.27	56.36	60.56	56.51
<i>Signal Element Relation</i>	36.52	36.68	38.45	39.04	35.38	47.88
<i>Lane Element Relation</i>	46.57	67.00	70.26	49.48	63.20	71.29
<i>Intention Judgment</i>	35.57	50.30	50.43	53.57	60.66	54.26
<i>Behavior Understanding</i>	11.50	41.56	44.36	47.49	45.47	50.28
<i>Ego Action Reasoning</i>	52.84	56.33	57.16	66.08	64.49	67.27
<i>Meta Decision-Making</i>	19.29	53.45	38.45	58.33	53.81	47.86
Overall	39.62	52.64	54.04	54.89	58.80	61.03

**VLADBench** for the small-scale VLMs and closed-source VLMs, are listed in Tab. 3. Besides, we present the results of large-scale VLMs in Tab. 4 for a thorough assessment. For comparison with the existing VLMs, we exclude our baseline model which serves for the following exploration. Then, we conduct the more experiments to explore the cognitive and reasoning interactions among the 5 key domains. The DS models trained on TKU data, GER data, TGG data, TAC data, EDP data, and the total data are compared with the base model for improvement visualization. Finally, we briefly discuss how the understanding of the five key domains by AD-specialized VLMs impacts the final trajectory prediction.

### 4.2.1. Evaluation on VLADBench

**Holistic Results.** The top score is held by the large-scale QW2.5-72B, which achieves 61.03, and followed by GPT-4o. For small-scale VLM models, LV-7B leads with a score of 53.63, which is 7.4 below the maximum. Across all the models, only one achieve more than 60 on our proposed benchmark, demonstrating the significant gap between current VLMs and human-level capabilities in real-world driving scenarios.

**Granular Results.** Through the results on 10 secondary aspects, the main findings are as follows:

- In TKU, *Road Traffic Signal* represents a fundamental knowledge of AD. Existing open-source models, with the

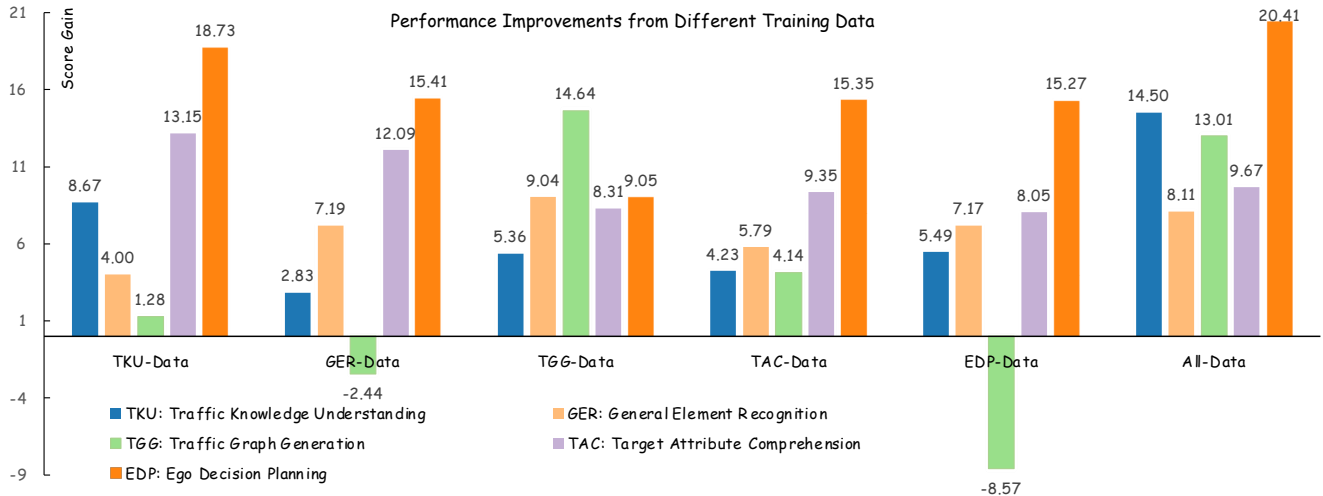


Figure 4. Score gains across five key domains. This chart illustrates the performance improvements of models trained on datasets categorized by the five key domains, evaluated on VLADBench, compared to the base model.

exception of large-scale QW series [4, 66], still remain a large room for performance improvement.

- On *Signal Element Relation* aspect, GPT-4o and QW2.5-72B [4] showcase the superiority, demonstrating the excellent spatial reasoning capability for AD scenarios.
- On *Intention Judgment* aspect, almost all the models exhibit unsatisfactory performance. However, QW-7B [3], trained without sequential data, achieves promising results, suggesting that VLMs can predict potential sequential events through training on non-sequential events.

**Large-scale vs. Small-scale VLMs.** To investigate how the performance of a general VLM varies with scalability, we present 6 large-scale models in Tab. 4. Firstly, the large-scale VLM models do not always surpass the small-scale models. Among the large-scale models, only four outperform the best small-scale VLM model LV-7B. Secondly, for a given model, scaling up the language model typically results in a overall performance improvement. However, this trend is not universally consistent in all aspects. For instance, with LV series, the larger model generally outperforms its smaller counterpart across most aspects, yet it underperforms in the important *intention judgment* and *meta decision-making* aspects.

**Domain-specific Results.** As training data from the AD domain is incorporated, the DS model exhibits outstanding performance in certain tasks, e.g., Dols-9B on vehicle cut-in task (86.70 ranked 1st), DriLM-4B on traffic light task (75.14 ranked 1st) and lane recognition task (73.44 ranked 1st). By comparing DS models and its base models, we can further observe that biased domain-specific data will lead to a loss of generalization ability in unseen tasks. DriLM-4B outperforms the base model (IVL2-4B) in fundamental traffic knowledge but performs significantly worse in the TAC domain. DS Data bias will influences the model capabilities, and single-direction optimization may lead to a loss of

generalization in other tasks even within the same domain.

#### 4.2.2. Interactions in Key Domains of VLADBench

As discussed above, biased domain-specific training data can enhance the performance in certain specialized areas of autonomous driving but may loss the generalization ability in tasks that require broader and more general knowledge. To deeply explore the interrelationships among the 5 key domains, we train the general IVL2-4B using different DS datasets (with generic data kept constant) and test the models on VLADBench. The score gains compared to IVL2-4B, are shown in Fig. 4. It can be concluded that:

- The contribution of each domain-specific dataset is not isolated, it also positively influences other domains. For example, TKU data boosts the EDP gain significantly, GER data benefits TAC gain, TGG data enhance the understanding of the traffic element.
- A synergy effect emerges when all datasets are combined for training. Compared to training on individual domains, joint training leads to higher performance gains across nearly all domains (except TGG), e.g., TKU gain improves by 67% (from 8.67 to 14.50).
- While the all-data model achieves notable improvements in the TGG domain, models trained solely on GER and EDP data exhibit negative gains in TGG, suggesting that the GER and EDP training datasets still carry a bias with respect to the TGG domain.
- Adjusting the data ratio across different domains may lead to improved training outcomes. For example, both TKU data and GER data contribute more to TAC gains than TAC data itself, suggesting that an appropriate balance of domain data could enhance overall performance..

#### 4.2.3. Contribution of Key Domains for Motion Planning

In VLADBench, tasks for comprehension dominate the evaluation. After assessing the 5 key domains for AD

Table 5. Motion planning performance in DriveLM-nuScenes [58] validation set.

Models	ST-P3 Metrics								UniAD Metrics							
	L2(m)↓				Collision(%)↓				L2(m)↓				Collision(%)↓			
	1s	2s	3s	Avg.	1s	2s	3s	Avg.	1s	2s	3s	Avg.	1s	2s	3s	Avg.
ST-P3	1.28	2.03	2.81	2.04	0.14	0.72	1.28	0.71	-	-	-	-	-	-	-	-
UniAD	-	-	-	-	-	-	-	-	0.47	1.80	3.73	3.00	0.13	0.53	1.50	0.72
TKU	0.30	0.67	1.15	0.70	0.13	0.25	0.82	0.40	0.43	1.28	2.41	1.37	0.13	0.51	2.53	1.06
GER	0.28	0.61	1.04	0.65	0.06	0.26	0.69	0.34	0.41	1.16	2.16	1.24	0.13	0.52	2.07	0.91
TGG	0.34	0.78	1.31	0.81	0.00	0.38	0.95	0.44	0.50	1.50	2.66	1.55	0.00	0.88	2.40	1.10
TAC	0.35	0.83	1.43	0.87	0.00	0.28	1.09	0.46	0.53	1.62	2.94	1.70	0.00	0.76	3.40	1.39
EDP	0.28	0.71	1.25	0.75	0.13	0.25	0.77	0.38	0.41	1.42	2.66	1.50	0.13	0.50	2.51	1.05
All	0.27	0.60	1.04	0.64	0.13	0.25	0.61	0.33	0.39	1.15	2.15	1.23	0.13	0.51	1.65	0.76

understanding tasks, we finally evaluate the contributions of 5 key domains for trajectory prediction. Note that the goal is not to pursue the state-of-the-art results. For training, we further construct about 4,000 samples from nuScenes [6], which includes scene analysis and trajectory points, and then train the model by incorporating individual DS data. The quantitative motion planning results are shown in Tab. 5. It can be observed that the GER is the most important domain for trajectory prediction, followed by EDP domain. The results from TKU domain are comparable to those from EDP domain, suggesting that the understanding of traffic knowledge plays a crucial role, which is a capability that traditional models are unable to achieve. Although the experimental results from TGG and TAC domains perform poorly in terms of L2 distance, they significantly reduce the collision rate in the short term. More details about the trajectory dataset and results from open-source and domain-specific models are presented in the supplement materials.

### 4.3. Further Analysis

**Bottlenecks in Traffic Graph Generation.** In TGG domain, analyzing the relations between elements presents significant challenges. As discussed above in Sec. 3.2, we incorporate descriptive guidance about the traffic elements as tips within the questions in TGG. Experiments after and before adding these tips showcase a up to 37% improvement rate in the *signal element relation* aspect and a up to 73% improvement rate in the *lane element relation* aspect, suggesting that embedding traffic-related knowledge can directly enhance knowledge graph construction. Nevertheless the final scores of *signal element relation* remain far from 60, indicating that spatial reasoning ability is still limited. The detailed improvement rates are listed in the supplement.

**The larger, the better?** OV-7B [31] and LV-7B [78] perform as the top models at the small scale. However, when the language model is scaled up to 72B, the vision encoder, SigLIP [76], remains unchanged, and the observed superiority no longer holds. OV-72B and LV-72B show only marginal improvement, and some aspects experiences a per-

formance decline. In contrast, IVL2-76B[78], with a significantly larger vision encoder (scaled from 300M to 6B parameters), achieved first place across two aspects. QW2-72B[66] and QW2.5-72B[4], featuring a larger vision encoder than OV[31] and LV[78] and employing a dynamic resolution mechanism to avoid visual information loss, approaches the performance of closed-source models, achieving a well-balanced performance across cognitive and reasoning tasks. These findings suggest that a large or specialized vision encoder may be more critical than merely scaling up the language model for AD.

## 5. Conclusion and Limitation

**Conclusion.** In this paper, we present a fine-grained benchmark for evaluating large vision-language models in autonomous driving. The proposed **VLADBench** covers 5 key domains, 11 aspects and 29 tasks, addressing critical gaps in current datasets, including coarse-grained categories, and limited analysis of dynamic elements and lack of data diversity. Extensive experiments on general and domain-specific models uncover the significant performance gaps across a wide range of tasks. Our in-depth experiments further reveals the interactions among the five key domains, and the individual contribution for motion planning performance.

**Limitations.** There are still several limitations: 1) The current benchmark focuses on evaluating the understanding and reasoning capabilities from the perspective view. Future research will incorporate multi-view inputs to further assess the 3D spatial perception capabilities of these models. 2) The training of domain-specific models in this paper is straightforward. Exploring the scalability of domain-specific models and optimizing data sampling strategies are the crucial directions for future researches.

## Acknowledgments

This work was supported in part by the Fundamental Research Funds for the Central Universities under Grant WK2100000059.

## References

- [1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 2
- [2] Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Shiori Sagawa, et al. Openflamingo: An open-source framework for training large autoregressive vision-language models. *arXiv preprint arXiv:2308.01390*, 2023. 1, 5
- [3] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 1(2):3, 2023. 6, 7
- [4] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 2, 6, 7, 8
- [5] Satantjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72, 2005. 3
- [6] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *Conference on Computer Vision and Pattern Recognition*, 2020. 3, 8
- [7] Xu Cao, Tong Zhou, Yunsheng Ma, Wenqian Ye, Can Cui, Kun Tang, Zhipeng Cao, Kaizhao Liang, Ziran Wang, James M Rehg, et al. Maplm: A real-world large-scale vision-language benchmark for map and traffic scene understanding. In *Conference on Computer Vision and Pattern Recognition*, pages 21819–21830, 2024. 6
- [8] Lin Chen, Jisong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. Sharegpt4v: Improving large multi-modal models with better captions. *arXiv preprint arXiv:2311.12793*, 2023. 2
- [9] Li Chen, Penghao Wu, Kashyap Chitta, Bernhard Jaeger, Andreas Geiger, and Hongyang Li. End-to-end autonomous driving: Challenges and frontiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 1, 3
- [10] Yinda Chen, Che Liu, Wei Huang, Sibao Cheng, Rossella Arcucci, and Zhiwei Xiong. Generative text-guided 3d vision-language pretraining for unified medical image segmentation. *arXiv preprint arXiv:2306.04811*, 2023. 1
- [11] Yuan Chen, Zi-han Ding, Ziqin Wang, Yan Wang, Lijun Zhang, and Si Liu. Asynchronous large language model enhanced planner for autonomous driving. In *European Conference on Computer Vision*, 2024. 3
- [12] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Conference on Computer Vision and Pattern Recognition*, 2024. 2, 6
- [13] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning, 2023. 2
- [14] Xinpeng Ding, Jianhua Han, Hang Xu, Wei Zhang, and Xiaomeng Li. Hilm-d: Towards high-resolution understanding in multimodal large language models for autonomous driving. *arXiv preprint arXiv:2309.05186*, 2023. 1
- [15] Xinpeng Ding, Jianhua Han, Hang Xu, Xiaodan Liang, Wei Zhang, and Xiaomeng Li. Holistic autonomous driving understanding by bird’s-eye-view injected multi-modal large models. In *Conference on Computer Vision and Pattern Recognition*, 2024. 1, 3
- [16] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, et al. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*, 2023. 3
- [17] Chaoyou Fu, Yuhang Dai, Yondong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. *arXiv preprint arXiv:2405.21075*, 2024. 3
- [18] Yunfei Guo, Fei Yin, Xiao-hui Li, Xudong Yan, Tao Xue, Shuqi Mei, and Cheng-Lin Liu. Visual traffic knowledge graph generation from scene images. In *International Conference on Computer Vision*, pages 21604–21613, 2023. 3, 6
- [19] Jianhua Han, Xiwen Liang, Hang Xu, Kai Chen, Lanqing Hong, Jiageng Mao, Chaoqiang Ye, Wei Zhang, Zhenguo Li, Xiaodan Liang, et al. Soda10m: A large-scale 2d self/semi-supervised object detection dataset for autonomous driving. *arXiv preprint arXiv:2106.11118*, 2021. 3
- [20] Wenyi Hong, Weihang Wang, Ming Ding, Wenmeng Yu, Qingsong Lv, Yan Wang, Yean Cheng, Shiyu Huang, Junhui Ji, Zhao Xue, et al. Cogvlm2: Visual language models for image and video understanding. *arXiv preprint arXiv:2408.16500*, 2024. 5
- [21] Sebastian Houben, Johannes Stallkamp, Jan Salmen, Marc Schlipsing, and Christian Igel. Detection of traffic signs in real-world images: The German Traffic Sign Detection Benchmark. In *International Joint Conference on Neural Networks*, 2013. 3
- [22] Yihan Hu, Jiazhi Yang, Li Chen, Keyu Li, Chonghao Sima, Xizhou Zhu, Siqi Chai, Senyao Du, Tianwei Lin, Wenhai Wang, Lewei Lu, Xiaosong Jia, Qiang Liu, Jifeng Dai, Yu Qiao, and Hongyang Li. Planning-oriented autonomous driving. In *Conference on Computer Vision and Pattern Recognition*, 2023. 3
- [23] Zhijian Huang, Chengjian Feng, Feng Yan, Baihui Xiao, Zequn Jie, Yujie Zhong, Xiaodan Liang, and Lin Ma. Drivemmm: All-in-one large multimodal model for autonomous driving. *arXiv preprint arXiv:2412.07689*, 2024. 6
- [24] Bo Jiang, Shaoyu Chen, Qing Xu, Bencheng Liao, Jiajie Chen, Helong Zhou, Qian Zhang, Wenyu Liu, Chang Huang,

- and Xinggang Wang. Vad: Vectorized scene representation for efficient autonomous driving. In *International Conference on Computer Vision*, 2023. 3
- [25] Bo Jiang, Shaoyu Chen, Bencheng Liao, Xingyu Zhang, Wei Yin, Qian Zhang, Chang Huang, Wenyu Liu, and Xinggang Wang. Senna: Bridging large vision-language models and end-to-end autonomous driving. *arXiv preprint arXiv:2410.22313*, 2024. 6
- [26] Siddharth Karamcheti, Suraj Nair, Annie S Chen, Thomas Kollar, Chelsea Finn, Dorsa Sadigh, and Percy Liang. Language-driven representation learning for robotics. *arXiv preprint arXiv:2302.12766*, 2023. 1
- [27] Jinkyu Kim, Anna Rohrbach, Trevor Darrell, John Canny, and Zeynep Akata. Textual explanations for self-driving vehicles. In *European conference on computer vision*, 2018. 6
- [28] Jinkyu Kim, Teruhisa Misu, Yi-Ting Chen, Ashish Tawari, and John Canny. Grounding human-to-vehicle advice for self-driving vehicles. In *Conference on Computer Vision and Pattern Recognition*, 2019. 3, 6
- [29] Iuliia Kotseruba, Amir Rasouli, and John K Tsotsos. Joint attention in autonomous driving (jaad). *arXiv preprint arXiv:1609.04741*, 2016. 3
- [30] Bohao Li, Yuying Ge, Yixiao Ge, Guangzhi Wang, Rui Wang, Ruimao Zhang, and Ying Shan. Seed-bench: Benchmarking multimodal large language models. In *Conference on Computer Vision and Pattern Recognition*, 2024. 3
- [31] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024. 6, 8
- [32] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, 2023. 2
- [33] Kaican Li, Kai Chen, Haoyu Wang, Lanqing Hong, Chaoqiang Ye, Jianhua Han, Yukuai Chen, Wei Zhang, Chunjing Xu, Dit-Yan Yeung, et al. Coda: A real-world road corner case dataset for object detection in autonomous driving. In *European Conference on Computer Vision*, 2022. 3, 6
- [34] KunChang Li, Yanan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. Videochat: Chat-centric video understanding. *arXiv preprint arXiv:2305.06355*, 2023. 1
- [35] Yanze Li, Wenhua Zhang, Kai Chen, Yanxin Liu, Pengxiang Li, Ruiyuan Gao, Lanqing Hong, Meng Tian, Xinhai Zhao, Zhenguo Li, et al. Automated evaluation of large vision-language models on self-driving corner cases. *arXiv preprint arXiv:2404.10595*, 2024. 3
- [36] Zhiqi Li, Zhiding Yu, Shiyi Lan, Jiahao Li, Jan Kautz, Tong Lu, and Jose M Alvarez. Is ego status all you need for open-loop end-to-end autonomous driving? In *Conference on Computer Vision and Pattern Recognition*, 2024. 3
- [37] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004. 3
- [38] Ji Lin, Hongxu Yin, Wei Ping, Pavlo Molchanov, Mohammad Shoeybi, and Song Han. Vila: On pre-training for visual language models. In *Conference on Computer Vision and Pattern Recognition*, 2024. 2
- [39] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2023. 1, 2
- [40] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, 2024. 2
- [41] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? In *European Conference on Computer Vision*, 2024. 3
- [42] Yuhang Lu, Yichen Yao, Jiadong Tu, Jiangnan Shao, Yuexin Ma, and Xinge Zhu. Can lvlms obtain a driver’s license? a benchmark towards reliable agi for autonomous driving. *arXiv preprint arXiv:2409.02914*, 2024. 3, 6
- [43] Yingzi Ma, Yulong Cao, Jiachen Sun, Marco Pavone, and Chaowei Xiao. Dolphins: Multimodal language model for driving. In *European Conference on Computer Vision*, 2024. 3, 6
- [44] Srikanth Malla, Chiho Choi, Isht Dwivedi, Joon Hee Choi, and Jiachen Li. Drama: Joint risk localization and captioning in driving. In *Winter Conference on Applications of Computer Vision*, 2023. 3, 6
- [45] Jiageng Mao, Minzhe Niu, Chenhan Jiang, Hanxue Liang, Jingheng Chen, Xiaodan Liang, Yamin Li, Chaoqiang Ye, Wei Zhang, Zhenguo Li, et al. One million scenes for autonomous driving: Once dataset. *arXiv preprint arXiv:2106.11037*, 2021. 3
- [46] Jiageng Mao, Yuxi Qian, Junjie Ye, Hang Zhao, and Yue Wang. Gpt-driver: Learning to drive with gpt. *arXiv preprint arXiv:2310.01415*, 2023. 1, 6
- [47] Ana-Maria Marcu, Long Chen, Jan Hünermann, Alice Karnsund, Benoit Hanotte, Prajwal Chidananda, Saurabh Nair, Vijay Badrinarayanan, Alex Kendall, Jamie Shotton, et al. Lingoqa: Visual question answering for autonomous driving. In *European Conference on Computer Vision*, 2024. 1, 3, 6
- [48] Michael Moor, Qian Huang, Shirley Wu, Michihiro Yasunaga, Yash Dalmia, Jure Leskovec, Cyril Zakka, Eduardo Pontes Reis, and Pranav Rajpurkar. Med-flamingo: a multimodal medical few-shot learner. In *Machine Learning for Health*, 2023. 1
- [49] Ming Nie, Renyuan Peng, Chunwei Wang, Xinyue Cai, Jianhua Han, Hang Xu, and Li Zhang. Reason2drive: Towards interpretable and chain-based reasoning for autonomous driving. In *European Conference on Computer Vision*, 2024. 1, 3
- [50] Chenbin Pan, Burhaneddin Yaman, Tommaso Nesti, Abhirup Mallik, Alessandro G Allievi, Senem Velipasalar, and Liu Ren. Vlp: Vision language planning for autonomous driving. In *Conference on Computer Vision and Pattern Recognition*, 2024. 3
- [51] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine

- translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 2002. 3
- [52] Chirag Parikh, Rohit Saluja, CV Jawahar, and Ravi Kiran Sarvadevabhatla. Idd-x: A multi-view dataset for ego-relative important object localization and explanation in dense and unstructured traffic. In *International Conference on Robotics and Automation*, 2024. 6
- [53] SungYeon Park, MinJae Lee, JiHyuk Kang, Hahyeon Choi, Yoonah Park, Juhwan Cho, Adam Lee, and DongKyu Kim. Vlaad: Vision and language assistant for autonomous driving. In *Winter Conference on Applications of Computer Vision Workshop*, 2024. 3
- [54] Tianwen Qian, Jingjing Chen, Linhai Zhuo, Yang Jiao, and Yu-Gang Jiang. Nuscenes-qa: A multi-modal visual question answering benchmark for autonomous driving scenario. In *AAAI Conference on Artificial Intelligence*, 2024. 3
- [55] Amir Rasouli, Iuliia Kotseruba, Toni Kunic, and John K Tsotsos. Pie: A large-scale dataset and models for pedestrian intention estimation and trajectory prediction. In *International Conference on Computer Vision*, 2019. 3
- [56] Enna Sachdeva, Nakul Agarwal, Suhas Chundi, Sean Roelofs, Jiachen Li, Mykel Kochenderfer, Chiho Choi, and Behzad Dariush. Rank2tell: A multimodal driving dataset for joint importance ranking and reasoning. In *Winter Conference on Applications of Computer Vision*, 2024. 3
- [57] Hao Shao, Yuxuan Hu, Letian Wang, Guanglu Song, Steven L Waslander, Yu Liu, and Hongsheng Li. Lmdrive: Closed-loop end-to-end driving with large language models. In *Conference on Computer Vision and Pattern Recognition*, 2024. 3
- [58] Chonghao Sima, Katrin Renz, Kashyap Chitta, Li Chen, Hanxue Zhang, Chengen Xie, Ping Luo, Andreas Geiger, and Hongyang Li. Drivelm: Driving with graph visual question answering. In *European Conference on Computer Vision*, 2024. 1, 3, 6, 8
- [59] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023. 6
- [60] Siyu Teng, Xuemin Hu, Peng Deng, Bai Li, Yuchen Li, Yunfeng Ai, Dongsheng Yang, Lingxi Li, Zhe Xuanyuan, Fenghua Zhu, et al. Motion planning for autonomous driving: The state of the art and future perspectives. *IEEE Transactions on Intelligent Vehicles*, 8(6):3692–3711, 2023. 1, 3
- [61] Xiaoyu Tian, Junru Gu, Bailin Li, Yicheng Liu, Yang Wang, Zhiyong Zhao, Kun Zhan, Peng Jia, Xianpeng Lang, and Hang Zhao. Drivevlm: The convergence of autonomous driving and large vision-language models. *arXiv preprint arXiv:2402.12289*, 2024. 1, 3
- [62] Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai Charitha Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, et al. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. *arXiv preprint arXiv:2406.16860*, 2024. 2
- [63] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 2
- [64] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Conference on Computer Vision and Pattern Recognition*, 2015. 3
- [65] Huijie Wang, Tianyu Li, Yang Li, Li Chen, Chonghao Sima, Zhenbo Liu, Bangjun Wang, Peijin Jia, Yuting Wang, Shengyin Jiang, et al. Openlane-v2: A topology reasoning benchmark for unified 3d hd mapping. *Advances in Neural Information Processing Systems*, 36:18873–18884, 2023. 6
- [66] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model’s perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 2, 6, 7, 8
- [67] Shihao Wang, Zhiding Yu, Xiaohui Jiang, Shiyi Lan, Min Shi, Nadine Chang, Jan Kautz, Ying Li, and Jose M Alvarez. Omnidrive: A holistic llm-agent framework for autonomous driving with 3d perception, reasoning and planning. *arXiv preprint arXiv:2405.01533*, 2024. 3
- [68] Benjamin Wilson, William Qi, Tanmay Agarwal, John Lambert, Jagjeet Singh, Siddhesh Khandelwal, Bowen Pan, Ratnesh Kumar, Andrew Hartnett, Jhony Kaesemodel Pontes, Deva Ramanan, Peter Carr, and James Hays. Argoverse 2: Next generation datasets for self-driving perception and forecasting. In *Neural Information Processing Systems Track on Datasets and Benchmarks*, 2021. 3
- [69] Yecheng Wu, Zhuoyang Zhang, Junyu Chen, Haotian Tang, Dacheng Li, Yunhao Fang, Ligeng Zhu, Enze Xie, Hongxu Yin, Li Yi, et al. Vila-u: a unified foundation model integrating visual understanding and generation. *arXiv preprint arXiv:2409.04429*, 2024. 5
- [70] Shaoyuan Xie, Lingdong Kong, Yuhao Dong, Chonghao Sima, Wenwei Zhang, Qi Alfred Chen, Ziwei Liu, and Liang Pan. Are vlms ready for autonomous driving? an empirical study from the reliability, data, and metric perspectives. *arXiv preprint arXiv:2501.04003*, 2025. 3
- [71] Zhenhua Xu, Yujia Zhang, Enze Xie, Zhen Zhao, Yong Guo, Kwan-Yee K Wong, Zhenguo Li, and Hengshuang Zhao. Drivegpt4: Interpretable end-to-end autonomous driving via large language model. *IEEE Robotics and Automation Letters*, 2024. 1, 3, 6
- [72] Fuzhao Xue, Yukang Chen, Dacheng Li, Qinghao Hu, Ligeng Zhu, Xiuyu Li, Yunhao Fang, Haotian Tang, Shang Yang, Zhijian Liu, et al. Longvila: Scaling long-context visual language models for long videos. *arXiv preprint arXiv:2408.10188*, 2024. 5
- [73] Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, et al. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*, 2024. 6
- [74] Jiabo Ye, Anwen Hu, Haiyang Xu, Qinghao Ye, Ming Yan, Yuhao Dan, Chenlin Zhao, Guohai Xu, Chenliang Li, Junfeng Tian, et al. mplug-docowl: Modularized multimodal large language model for document understanding. *arXiv preprint arXiv:2307.02499*, 2023. 1

- [75] Jiang-Tian Zhai, Ze Feng, Jinhao Du, Yongqiang Mao, Jiang-Jiang Liu, Zichang Tan, Yifu Zhang, Xiaoqing Ye, and Jingdong Wang. Rethinking the open-loop evaluation of end-to-end autonomous driving in nuscenes. *arXiv preprint arXiv:2305.10430*, 2023. [3](#)
- [76] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *International Conference on Computer Vision*, pages 11975–11986, 2023. [8](#)
- [77] Pan Zhang, Xiaoyi Dong, Bin Wang, Yuhang Cao, Chao Xu, Linke Ouyang, Zhiyuan Zhao, Haodong Duan, Songyang Zhang, Shuangrui Ding, et al. Internlm-xcomposer: A vision-language large model for advanced text-image comprehension and composition. *arXiv preprint arXiv:2309.15112*, 2023. [2](#), [5](#)
- [78] Yuanhan Zhang, Bo Li, haotian Liu, Yong jae Lee, Liangke Gui, Di Fu, Jiashi Feng, Ziwei Liu, and Chunyuan Li. Llava-next: A strong zero-shot video understanding model. 2024. [6](#), [8](#)
- [79] Yi-Fan Zhang, Huanyu Zhang, Haochen Tian, Chaoyou Fu, Shuangqing Zhang, Junfei Wu, Feng Li, Kun Wang, Qing-song Wen, Zhang Zhang, et al. Mme-realworld: Could your multimodal llm challenge high-resolution real-world scenarios that are difficult for humans? *arXiv preprint arXiv:2408.13257*, 2024. [3](#)
- [80] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023. [2](#)
- [81] Brianna Zitkovich, Tianhe Yu, Sichun Xu, Peng Xu, Ted Xiao, Fei Xia, Jialin Wu, Paul Wohlhart, Stefan Welker, Ayzaan Wahid, et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. In *Conference on Robot Learning*, 2023. [1](#)