This ICCV paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Passing the Driving Knowledge Test

Maolin Wei[1*] Wanzhou Liu[2*] Eshed Ohn-Bar[1]
[1]Boston University [2]Washington University in St. Louis

Abstract

*If a Large Language Model (LLM) were to take a driving knowledge test today, would it pass? Beyond standard spatial and visual question-answering (QA) tasks on current autonomous driving benchmarks, driving knowledge tests require a complete understanding of all traffic rules, signage, and right-of-way principles. To pass this test, human drivers must discern various edge cases that rarely appear in real-world datasets. In this work, we present **DriveQA**, an extensive open-source text and vision-based benchmark that exhaustively covers traffic regulations and scenarios. Through our experiments using DriveQA, we show that (1) state-of-the-art LLMs and Multimodal LLMs (MLLMs) perform well on basic traffic rules but exhibit significant weaknesses in numerical reasoning and complex right-of-way scenarios, traffic sign variations, and spatial layouts, (2) fine-tuning on DriveQA improves accuracy across multiple categories, particularly in regulatory sign recognition and intersection decision-making, (3) controlled variations in DriveQA-V provide insights into model sensitivity to environmental factors such as lighting, perspective, distance, and weather conditions, and (4) pretraining on DriveQA enhances downstream driving task performance, leading to improved results on real-world datasets such as nuScenes and BDD, while also demonstrating that models can internalize text and synthetic traffic knowledge to generalize effectively across downstream QA tasks.*

1. Introduction

Safe navigation in traffic requires not only recognizing and interpreting visual information but also reasoning over traffic rules and making decisions that align with regulations. To ensure drivers develop these critical skills, before receiving their license they must first pass a written knowledge test—a structured (multiple choice questions) assessment designed to evaluate precise understanding of traffic laws, right-of-way rules, and complex driving scenarios [32, 56].

Driving tests are not merely procedural; they assess a driver's ability to apply reasoning across a wide range of traffic conditions. While primarily textual, these tests may also include graphical illustrations to ground questions in real-world scenarios. Recent advances in Multimodal Large Language Models (MLLMs) [3, 26, 47, 75, 97] as general-purpose reasoning models provide an opportunity to explore a key question: how well do current vision-and-language models perform when faced with the same driving knowledge assessments? Even without targeted fine-tuning, MLLMs may inherit some traffic rule knowledge from their pretraining data (however, our findings indicate that both such knowledge and associated reasoning capabilities remain limited).

Researchers have been increasingly integrating MLLMs into autonomous driving systems [11, 14, 28, 39, 42, 51, 53, 54, 70, 77, 84, 89, 95, 98]. However, while these models are often tested on perception-focused benchmarks that em-

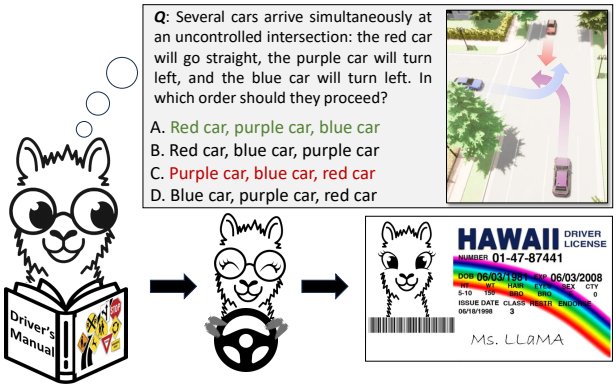

Figure 1. **Can LLMs Pass a Driving Knowledge Test?** We introduce a comprehensive multimodal dataset to evaluate the traffic rule-following capabilities of MLLMs. While most question answering (QA) benchmarks in autonomous driving focus on spatial understanding and common planning tasks, our **DriveQA** dataset assesses broad driving knowledge. The challenging benchmark comprises text-based questions derived from various U.S. state driving manuals, as well as visual tasks for traffic sign recognition and right-of-way judgment. We evaluate both text-only and image-text QA using synthetic images (varying perspectives, weather, time of day, and sign types) while showing transferability and generalization to downstream real-world driving tasks. Ths figure shows one example for a right-of-way question, a category where models frequently struggle. Incorrect responses are highlighted in red and correct answers in green.

*Equally contributed.

Table 1. **A Multimodal Dataset Emphasizing Traffic Rules.** The table compares existing benchmarks in terms of: the total number of images (**#Images**), the number of QA pairs (**#QA Pairs**), the method of annotation (**Annotations**, A+M means semi-automatic labeling), environment settings (**Settings**, including camera perspective of forward-**Fwd**, oblique-**Obl**, and top-down-**Top** views, weather, and time of day conditions), explanations for each question's answer (**Explanations**), and traffic rule reasoning (**Traffic Rules**), which is our focus.

Benchmarks	Image Source	#Images	#QA Pairs	Annotations	Settings			Explanations	Traffic Rules
					Perspective	Weather	Time of Day		
EQA-v1 [19]	House3D [85]	767	5,281	A	Fwd, Obl	✗	✗	✗	✗
OpenEQA [52]	OpenEQA	180	1,600	M	Fwd, Obl	✗	✗	✗	✗
SpatialVLM [9]	Internet	10M	2B	A	Fwd, Obl	✗	✗	✗	✗
NuScenes-QA [63]	nuScenes [6]	34K	450K	A	Fwd	✓	✓	✗	✗
DriveLM-nuScenes [72]	nuScenes [6]	4,871	443K	A+M	Fwd	✓	✓	✗	✗
DriveLM-CARLA [72]	CARLA [22]	64,285	1,566K	A	Fwd	✓	✓	✗	✗
DriveBench [86]	nuScenes [6]	19,200	20,498	A+M	Fwd	✓	✓	✗	✗
LingoQA [55]	LingoQA	28k	419.9K	A+M	Fwd	✓	✓	✗	✗
DriveQA-V (ours)	CARLA [22], Mapillary [57]	68K	448K	A+M	Fwd, Obl, Top	✓	✓	✓	✓
DriveQA-T (ours)	-	-	26K	A+M	-	✓	✓	✓	✓

phasize spatial awareness and standard planning tasks (e.g., lane keeping, collision avoidance [44, 78, 87]), they are rarely evaluated for their ability to understand and comply with diverse traffic regulations, such as reasoning about traffic rules, reacting safely to no-entry signs, or maintaining speed limit. While most existing datasets narrowly focus on perception and basic trajectory planning, driving knowledge tests are designed to assess a broad spectrum of all regulations, including rare traffic signs, difficult right-of-way cases, and edge-case rules that are essential for safe navigation but seldom appear in collected driving data. This highlights a critical gap in evaluating AI systems: while they may perform well in current benchmarks, their ability to reason over long-tail traffic rules and regulatory compliance remains understudied. There is also substantial anecdotal evidence suggesting that current commercial systems, e.g., Tesla's Full Self-Driving [24, 25, 36, 79], often struggle with interpreting traffic rules.

To address this gap and enhance the evaluation of reasoning capabilities in both LLMs and MLLMs, we introduce a novel driving knowledge benchmark, **DriveQA**. Our dataset includes both text-only question-answers (QA) and aligned image-text (VQA) pairs. Thus, we enable the first thorough evaluation of vision-and-language model performance across broad driving tasks, from basic regulatory questions and signs to complex multimodal reasoning tasks. Our **contributions** are summarized as follows:

- We introduce **DriveQA**, a large-scale benchmark featuring both text-based (**DriveQA-T**) and vision-based (**DriveQA-V**) driving knowledge assessments. To ensure broad coverage of traffic regulations, right-of-way rules, and rare driving scenarios, we leverage synthetic procedural data generation with comprehensive traffic reasoning, controlled variations (e.g., sign placement and weather), and new 3D sign assets integrated into CARLA [21], as well as manually annotated real-world data from Mapillary [57]. DriveQA covers 19 question categories, 220 traffic signs, and 474K samples.
- We *benchmark* state-of-the-art LLMs and MLLMs on DriveQA to uncover that while these models perform well on basic traffic rules, they struggle with numerical precision, right-of-way reasoning, spatial awareness, and environmental sensitivity (e.g., time-of-day, perspective, and geometric layouts). Our findings suggest that MLLMs inherit limited traffic knowledge from pretraining and require fine-tuning for our task.
- We demonstrate the effectiveness of DriveQA *pretraining*; models trained on our text and purely synthetic data demonstrate improved performance across various real-world driving tasks [87, 88]. We show that pretraining on DriveQA improves the performance on both trajectory prediction and driving action reasoning tasks. This highlights its role in evaluating and enhancing multimodal reasoning, and as a step toward bridging theory and practice in embodied AI systems that can learn to make decisions in the real-world based on text or synthetic data.

2. Related Work

Based on our survey of MLLM-based studies and VQA benchmarks for autonomous driving below, we find prior work rarely addressed traffic rules, signage, and right-of-way principles within their driving knowledge assessments. Relevant related benchmarks are compared in Table 1.

Multimodal Large Language Models: Our study diagnoses multimodal reasoning capabilities in MLLMs [2, 17, 38, 61, 78, 80, 95, 97]. A typical MLLM architecture comprises three main modules: a pre-trained modality encoder, a pre-trained language model, and a modality projector that aligns them. The modality encoder processes non-textual inputs, such as images, transforming them into representations compatible with the language models. Vision Transformer (ViT) [23] is widely used to extract image features. For example, CLIP [65] leverages ViT as its visual encoder to transform images into feature representations that align with text through extensive pre-training on large-scale

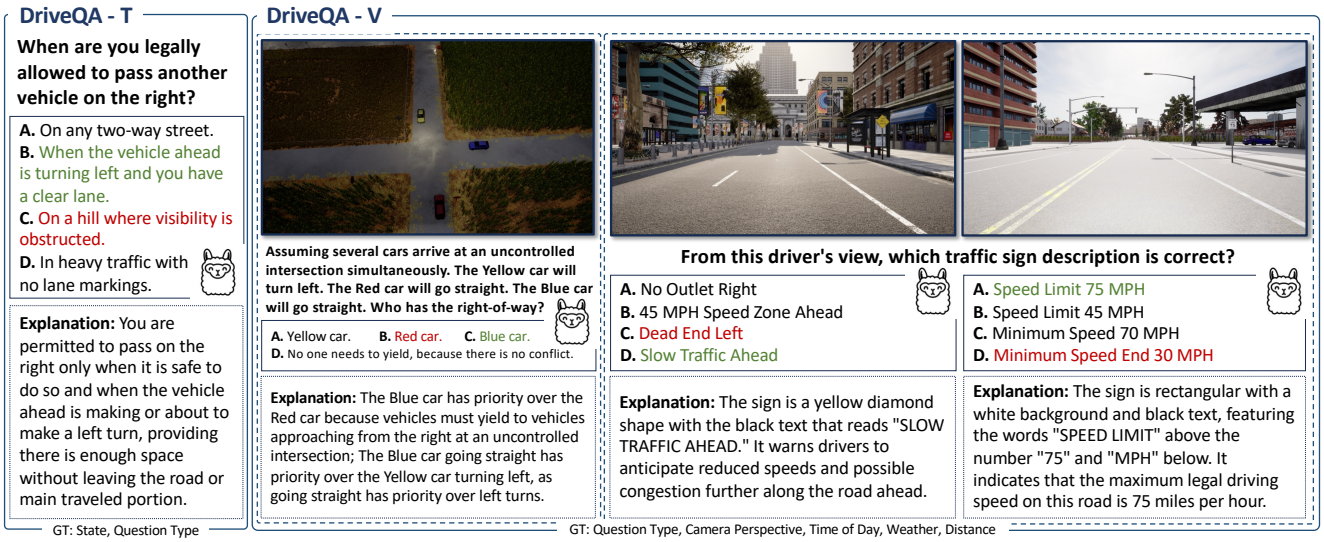

Figure 2. **Example Questions and Answers of DriveQA Dataset.** We introduce a text and vision-based benchmark for extensively validating driving knowledge with question type categorization, answer explanation, and environmental information ground truth (GT).

image-text pairs. The modality projector aligns encoder outputs with the language model, enabling integration of modality data with text. A common approach is to use a set of learnable query tokens to extract information in a query-driven manner [7], which has been employed by a variety of models [10, 13, 18, 40, 43, 90–92]. Additionally, methods may design MLPs to transform the high-dimensional input features into a unified representation [2, 50, 62, 73]. Our systematic study controlling for variations in QA category and image factors reveals limitations of current alignment mechanisms in supporting multimodal or spatial reasoning.

MLLM-based Driving Agents: While recent advancements have applied MLLMs to autonomous driving tasks, most focus on leveraging reasoning and language understanding capabilities to improve driving decisions in narrow tasks [12, 15, 16, 28, 53, 70, 82, 89, 98]. For instance, several vision-and-language agents for motion planning and decision-making have been proposed and evaluated on datasets such as nuScenes [4, 6, 28, 42, 51, 53, 54, 67, 78, 89]. The key hypothesis in such studies is that MLLMs can inherit general-purpose reasoning and knowledge from their pretraining; however, our findings suggest that while they may grasp basic traffic concepts, their ability to apply traffic reasoning in driving-specific scenarios remains limited. Moreover, these works have not explicitly addressed MLLMs' ability to comprehend diverse traffic rules and regulations-a critical requirement for safe driving.

Datasets for Autonomous Driving: Several real-world, synthetic, and VQA benchmarks for autonomous driving are currently being used to evaluate driving models, including KITTI [30, 45], Waymo Open [74], Argoverse [8, 83], and nuScenes [6]. However, few incorporate more than a handful of traffic rules, e.g., researchers may evaluate collision on nuScenes [6, 20, 34, 78, 93, 94, 99], yet lack coverage and exclude explicitly evaluating for traffic signs or right-of-way reasoning. Crowdsourced benchmarks such as Mapillary [57], which we augment with VQA annotations, are broad but still lack in long-tail events, motivating the use of synthetic benchmarks. Yet, prior simulation-based studies (e.g., [1, 21, 27, 35, 66, 68, 71, 96]) have only accounted for a handful of potential regulatory and safety violations. For instance, while CARLA [21] enables controllable and diverse data generation (e.g., perspectives, scenarios, weather), most traffic signs are missing in CARLA, a limitation addressed by our work. The development of MLLMs and their applications in autonomous driving lead to the emergence of driving vision-language datasets [5, 55, 64, 69, 72, 78, 80] specifically designed to support vision understanding and reasoning in complex driving scenarios. However, here as well existing efforts focus on scene understanding, perception and basic planning (i.e., collision avoidance, intersection boundary [44]), neglecting reasoning about traffic rules and regulations (i.e., reacting safely to no-entry signs, maintaining speed limit, etc.) which is a foundational driving test for humans.

3. A Multimodal Driver Knowledge Test

In this section, we outline our scalable data collection and annotation process. Our dataset consists of QA pairs that cover essential aspects of real-world driving knowledge. As illustrated in Fig. 2, our dataset comprises two tasks: DriveQA-T, which consists of text-based QA pairs on general driving rules, and DriveQA-V, focusing on visual (image-based) QA related to traffic sign comprehension and right-of-way scenarios. We adhere as closely as possible to standard driving knowledge tests to ensure meaningful

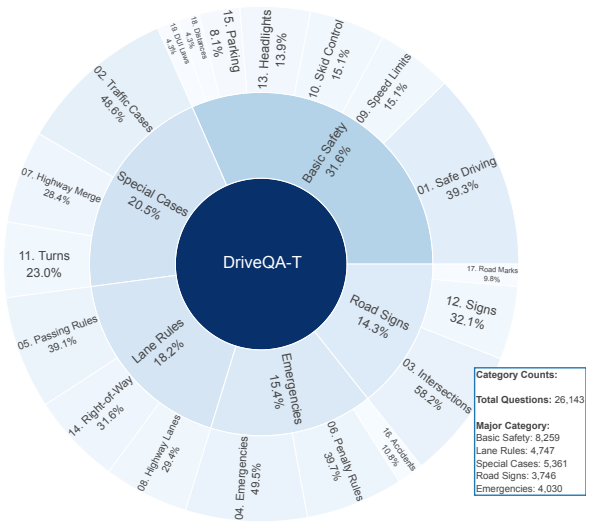

Figure 3. **Distribution of Question Type in DriveQA-T.** The benchmark covers five key domains and 19 sub-class types.

comparisons to human performance on these assessments, and generate a diverse set of multiple-choice questions. We note that there are commercial driver knowledge tests available [56, 76], however these are closed-source. To ensure in-depth analysis, we further provide reasoning for ground-truth answers on both tasks. This design is intended to provide a holistic and systematic analysis of both LLMs and MLLMs in decision-oriented tasks. Ultimately, our overarching goal is to enable novel mechanisms to teach MLLMs real-world tasks, e.g., through text descriptions or synthetic examples.

Text-based QA Dataset—DriveQA-T: Our DriveQA-T dataset contains a total of 26K QA pairs covering different general driving topics, including traffic lights, traffic signs, parking, regulation, and symbols (see our supplementary for full details on the categories). Each QA pair contains an explanation for the correct answer, which can be used to evaluate the reasoning capabilities of LLMs. To curate the QA pairs, we first gathered 51 official driver's handbooks from all 50 US states plus DC. Although our data set is US-centric, it can inform the construction of additional international datasets in the future. We build DriveQA-T in three steps. First, we generate questions automatically by prompting GPT-4o [59, 98] with the driver's handbooks as context, and then conduct manual quality verification based on the driver's handbooks. Quality checks were performed in rounds, where each verifier went through questions, and then ambiguous or inconsistent cases were discarded. Additional details about this process can be found in the supplementary. We note that humans, once trained, can obtain 100% on our benchmark. We categorized the text data into 19 classes, grouped into five main categories, as shown in Fig. 3. A summarized description of the dataset is depicted

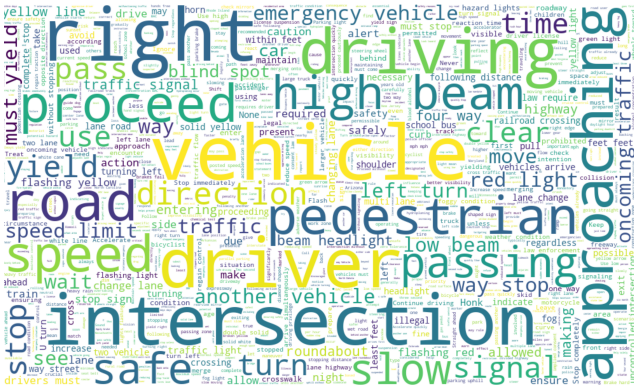

Figure 4. **Word Cloud of Questions in DriveQA.** The figure statistically summarizes the language terms in the introduced DriveQA benchmark.

in Fig. 4, showing a focus on traffic participants and intersections (e.g., right-of-way, yielding behaviors).

Multimodal Extension With DriveQA-V: Driver knowledge tests [56, 76] are primarily text-based, e.g., with a full description of objects and spatial layout information in text. However, certain questions particularly related to *traffic signs* and *right-of-way*, test understanding through graphical illustrations accompanying text information. DriveQA-V focuses on these two types of questions. To ensure comprehensive coverage through procedural variations (e.g., camera perspectives, time of day, weather, distance), images are collected with the open-source Unreal Engine-based CARLA simulator [21]. However, since CARLA was not originally designed with extensive traffic rule knowledge, e.g., traffic signs, we augment the simulation with additional 3D assets and automatic traffic rule scripts. Due to procedural and synthetic generation, in addition to aligned text-image VQA pairs, we are able to collect full state information, such as camera perspective, distance from ego-vehicle, and sign type. Specifically, we insert 220 US-based traffic sign models into the map, simulator, and spawn an ego vehicle to collect sensor readings. For right-of-way questions, we identify intersections in the CARLA maps and randomly spawn vehicles on each side of the intersection. Each vehicle varies in color to facilitate identification in the questions.

4. Method

In this section, we describe our approach to evaluating models on our proposed dataset. The methodology includes question-type classification, model evaluation using Chain of Thought (CoT) [81] reasoning, Retrieval-Augmented Generation (RAG) [41] techniques, and model fine-tuning on the benchmark.

Question-Type Classification: To precisely assess model performance across specific traffic rule categories, we di-

vide the DriveQA-T dataset questions into types. This enables us to assess performance on specific traffic rule categories, thereby providing a nuanced understanding of how well they generalize across various traffic contexts. Specifically, we apply hierarchical clustering [58] to organize questions into semantically coherent groups, ensuring that similar questions are grouped together based on their thematic content. We begin by generating embeddings for each question using BERT [37], which effectively captures the semantic nuances of each question and represents them in a high-dimensional embedding space. By applying hierarchical clustering to these embeddings, we identify clusters that correspond to distinct traffic rule topics, such as traffic signals, speed limits, parking regulations, etc. To interpret and label each cluster, we use KeyBERT [31] to extract semantic keywords for each group, combined with sample questions from each cluster, finally we assign descriptive types to the clusters. In DriveQA-V, we assign types manually (see supplementary for more details).

Fine-Tuning: Off-the-Shelf models were trained on open web data, thus having potential access to driver handbooks and tests. To further investigate the role of training data for our task, we also fine-tune models on our dataset. We find this to enhance, but not fully address, models' ability to handle the specific complexities of traffic scenarios. We employ LoRA [33], which reduces the number of trainable parameters by introducing low-rank updates to the weight matrices in transformer layers, allowing efficient fine-tuning without requiring extensive computational resources.

CoT and RAG: We employ CoT reasoning and RAG-based context in our evaluation. CoT reasoning guides the LLMs and MLLMs through each reasoning step in a logical progression, which allows us to test their capacity for logical consistency, especially in multi-vehicle or rule-based scenarios. We also evaluate the produced reasoning, e.g., to ensure correct answers are selected for the correct reasons. For RAG, we construct a retrieval corpus derived from the official driver's handbooks of all 50 U.S. states and DC. This corpus serves as a reliable, contextually relevant reference to provide the models with related context when answering questions. By retrieving it for each question, RAG-based context grounds the model's responses in actual regulations, aiming to enhance both the accuracy and contextual relevance of answers.

5. Experiments

5.1. Setup

We evaluate our dataset on various MLLMs. For each model type, we consider both open-source and closed-source variants, applying CoT and RAG strategies to structure the input prompts. Our evaluation is based not only on testing the original capabilities of each off-the-shelf model but also on a comprehensive analysis of their performance after fine-tuning the open-source checkpoint on our dataset.

Prompt Structure: We designed four prompt structures to explore model performance under varying levels of reasoning and contextual support. Beginning with a basic prompt, we tested standard question-answering without additional guidance. Building on this, we introduced a CoT prompt to encourage step-by-step reasoning, aiming to enhance answer consistency in complex scenarios. To further improve contextual relevance, we combined CoT with RAG-based context by retrieving pertinent information from drivers' handbooks, thereby grounding the responses in real-world regulations. Finally, we assessed the impact of RAG-based context alone, where we provided retrieved contextual information without step-by-step reasoning. These four prompts allowed us to examine the models' capabilities in integrating both reasoning and factual support effectively.

Metrics: To comprehensively evaluate our model's performance on both the DriveQA-T and DriveQA-V datasets, we use accuracy as the primary metric, reflecting the model's ability to correctly answer a wide range of driving-related questions across textual and visual domains. For the DriveQA-T dataset, we place an additional emphasis on reasoning capability, as each question includes an accompanying explanation. To measure the relevance of the model's reasoning, we employ BLEU-4 [60] and ROUGE-L [46], providing insights into the model's ability to generate responses that are not only accurate but also demonstrate high-quality reasoning aligned with expected standards.

5.2. Results

Performance of LLMs on DriveQA-T: Table 2 presents the performance of various models on our DriveQA-T dataset. Phi-3.5-mini and Gemma-2 (9B) generally perform better across most categories than other models, demonstrating their ability to comprehend driving rules. Observably, models with CoT reasoning and RAG-based context tend to achieve higher accuracy, suggesting that these enhancements contribute to performance improvements. This trend highlights the importance of advanced reasoning and contextual retrieval for complex, rule-based tasks. While certain models show promising results in accurately interpreting and following traffic regulations, consistent performance across diverse driving-related categories may still require further refinement.

As shown in Table 2, all models exhibit a significant improvement in overall accuracy after fine-tuning. However, they still struggle with numerical questions, such as those in the "`Limits`" and "`Alcohol`" categories. This difficulty suggests that models may lack the precise numerical reasoning capabilities needed to respond accurately to questions involving specific values or quantitative thresh-

Table 2. **Challenging Categories on DriveQA-T**. We show the results of most difficult 3 types: Limits: Speed and Distance Limits, Parking: Parking and Wheel Positioning, Intersection: Right-of-Way and Lane Selection. The Average is the summary based on all 19 types of questions. We denote with **green the top method**, and **light green second best**.

Models	Size	CoT	RAG	Finetune	Limits	Parking	Intersection	Average
Gemma-2 [75]	2B				42.15	35.64	27.88	44.15
		✓			42.98	42.57	34.51	52.77
		✓	✓		58.68	47.52	55.75	56.62
		✓	✓	✓	62.40	61.39	85.84	72.01
Gemma-2 [75]	9B				57.85	54.46	58.41	71.00
		✓			59.50	58.42	62.83	72.20
		✓	✓		64.88	68.32	77.88	76.91
		✓	✓	✓	72.31	88.12	91.15	87.28
Llama-3.1 [26]	8B				53.72	37.62	48.23	55.89
		✓			55.37	38.61	65.93	56.22
		✓	✓		55.37	46.53	68.58	60.79
		✓	✓	✓	72.73	86.14	91.59	87.62
Llama-3.2 [26]	3B				36.78	35.64	42.92	50.93
		✓			48.35	26.73	49.56	48.92
		✓	✓		61.16	53.47	61.50	64.19
		✓	✓	✓	69.42	75.25	85.84	82.82
Phi-3.5-mini [3]	3.8B				49.17	48.51	79.65	69.79
		✓			55.79	45.54	79.65	71.14
		✓	✓		63.22	62.38	84.96	77.30
		✓	✓	✓	66.94	65.35	87.17	81.08
GPT-4o [59]	-	✓	✓		76.72	93.75	97.27	91.96

Table 3. **Performance of CoT Reasoning on DriveQA-T.** The evaluation includes both off-the-shelf and fine-tuned models under two settings of with and without RAG.

Models	Size	BLEU-4		ROUGE-L	
		w/o RAG	w/ RAG	w/o RAG	w/ RAG
Off-The-Shelf Models					
Gemma-2 [75]	2B	0.1098	0.1704	0.2920	0.3387
Gemma-2 [75]	9B	0.3234	0.3116	0.4295	0.4276
Llama-3.1 [26]	8B	0.2573	0.2619	0.3270	0.3317
Llama-3.2 [26]	3B	0.2258	0.3140	0.3348	0.4024
Phi-3.5-mini [3]	3.8B	0.2437	0.2574	0.3616	0.3996
GPT-4o [59]	-	0.3905	0.3989	0.5354	0.5393
Finetuned Models					
Gemma-2 [75]	2B	0.3623	0.2934	0.5058	0.4458
Gemma-2 [75]	9B	0.4112	0.4105	0.5420	0.5528
Llama-3.1 [26]	8B	0.3042	0.2946	0.4749	0.4750
Llama-3.2 [26]	3B	0.2131	0.1916	0.3853	0.3570
Phi-3.5-mini [3]	3.8B	0.2362	0.1891	0.4073	0.3476

olds, which are critical in understanding speed limits, alcohol levels, and other regulatory metrics. Furthermore, for certain decision-making-focused categories, including "`Passing`", "`Signs`" and "`Turning`", most models achieve only slightly above accuracy of 80%. These categories are crucial for safe driving in practical conditions, highlighting the models' continuous shortcomings in handling nuanced, context-dependent traffic rules despite fine-tuning improvements.

CoT Reasoning of LLMs on DriveQA-T: Table 3 shows the evaluation results of CoT reasoning on the DriveQA-T dataset. Most models show improvements when using RAG-based context. Specifically, GPT-4o achieves the highest BLEU-4 and ROUGE-L scores among the off-the-shelf models, reaching a BLEU-4 score of 0.3989 and a ROUGE-L score of 0.5393 with RAG-based context. After fine-tuning, Gemma-2 (9B) surpasses GPT-4o in both BLEU-4 and ROUGE-L scores, demonstrating the effectiveness of fine-tuning in adapting the model specifically to traffic rules and enabling it to provide more accurate, context-specific explanations. However, these scores still fall short of what would be considered high-quality for generating fully robust and exhaustive explanations, indicating that the models are not yet capable of consistently producing complete and nuanced responses. Furthermore, the lower scores of Llama-3.2 and Phi-3.5-mini after fine-tuning suggest potential issues. One possible reason for this decline is overfitting the fine-tuning dataset, which may cause the models to become too specialized and lose some of their generalization capabilities. This overfitting can result in explanations that are overly tailored to specific training examples, reducing the models' ability to produce flexible, broadly applicable responses. Additionally, fine-tuning may interfere with the effectiveness of RAG-based retrieval, leading to less relevant contextual information and, consequently, lower alignment with ground-truth explanations. These factors highlight the challenges of balancing specificity and generalization in fine-tuning for complex, rule-based tasks.

Performance of MLLMs on DriveQA-V: Table 4 presents the accuracy of MLLM models on DriveQA-V, which assesses model performance across intersection types and traffic sign categories. The dataset divides intersections into 4 different categories based on the intersection types and camera perspective, and 4 different categories of signs based on most states' driver handbooks. Among the off-the-shelf models, GPT-4o achieves the highest accuracy in all intersection and sign categories, with a particularly strong performance in the sign types (around 94%). This suggests that GPT-4o possesses a deep understanding of signs. However, for intersection-based categories, the performance remains relatively low, with the highest off-the-shelf accuracy of 60.36% in the "`T-Top`" category. Most models except GPT-4o perform below random guess level (25%) in several categories due to bias [2]. This indicates that off-the-shelf models struggle to fully understand and apply traffic rules in intersection scenarios, which often require more complex visual-spatial reasoning. Additionally, Fine-tuning significantly enhances model performance across all categories. All models achieve notable improvements after fine-tuning, which demonstrates that fine-tuning effectively adapts MLLMs to handle the visual-spatial and contextual nuances for the accurate understanding of both right-of-way rules and traffic signs.

Despite these gains, there remain limitations. Both LLaVA-1.5 and VILA-1.5, even after fine-tuning, achieve only moderate accuracy in intersection categories, with par-

Table 4. **Summarized Results on DriveQA-V.** We show model performance (accuracy %) for VQA. The dataset is divided into two main categories: intersections and signs (categorized into camera perspective and type).

Models	Size	DriveQA-V (Inters.)				DriveQA-V (Signs)				Average
		T-Front	T-Top	Cross-Front	Cross-Top	Regulatory	Warning	Guide	Temporary Control	
Off-The-Shelf Models										
Mini-InternVL [29]	2B	27.83	24.83	26.00	25.65	64.06	55.34	65.82	45.04	41.82
LLaVA-1.5 [48]	7B	23.30	23.10	24.96	23.24	23.51	26.61	22.31	21.10	23.52
LLaVA-1.6-mistral [49]	7B	18.77	19.66	30.99	30.47	42.58	43.01	52.75	37.50	34.47
VILA-1.5 [47]	8B	15.53	16.86	15.69	20.35	25.32	23.33	27.78	21.46	20.79
GPT-4o [59]	-	55.09	60.36	50.52	59.14	93.75	94.02	95.11	94.35	75.29
Finetuned Models										
Mini-InternVL [29]	2B	86.73	82.07	74.33	76.01	93.79	92.19	91.08	96.51	86.59
LLAVA-1.5 [48]	7B	64.18	70.57	54.77	56.52	72.22	73.00	76.82	89.27	69.67
LLaVA-1.6-mistral [49]	7B	86.08	85.52	74.38	74.53	82.05	84.10	88.11	94.49	83.66
VILA-1.5 [47]	8B	47.67	52.27	55.60	57.26	87.10	83.14	91.46	95.33	71.23

Table 5. **10 Most Difficult Sign Types in DriveQA-V.** We calculate the lowest accuracy over all the models' performance based on different sign types. Most challenging cases belong to regulatory and warning signs.

Model	Size	Playground	Trauma Center	Golf Carts	Ground Clearance	No Stopping	No Parking	Push Button	Weekday Only	Fire Truck	Tractor Crossing
Off-The-Shelf Models											
Mini-InternVL [29]	2B	0.00	27.78	14.81	15.00	42.10	48.14	0.00	8.70	0.00	20.00
LLaVA-1.5 [48]	7B	5.26	2.38	5.43	11.76	16.30	10.42	12.50	20.83	20.45	21.59
LLaVA-1.6-mistral [49]	7B	0.00	16.67	25.93	25.00	5.00	22.22	48.00	17.39	0.00	24.00
VILA-1.5 [47]	8B	1.32	2.38	0.00	8.82	8.70	5.21	2.78	19.79	3.41	31.82
Finetuned Models											
Mini-InternVL [29]	2B	88.46	94.44	88.88	95.00	100.00	90.91	96.00	100.00	92.86	80.00
LLaVA-1.5 [48]	7B	73.68	85.71	61.96	64.71	65.22	61.46	75.00	37.50	59.09	57.95
LLaVA-1.6-mistral [49]	7B	80.77	83.33	74.07	85.00	85.19	77.27	92.00	100.00	92.86	92.00
VILA-1.5 [47]	8B	68.42	59.52	83.70	66.18	65.22	79.17	66.67	80.21	80.68	52.27

ticularly lower performance on first-person perspective images. This suggests that the models still struggle with complex, multi-vehicle intersection scenarios, where perspective and spatial relationships are critical. For the traffic signs recognition task, We can observe the best training performance in the Guide Signs and Temporary Traffic Control categories. This is because guide signs typically feature simpler images with blue backgrounds, while temporary traffic control signs have distinct orange backgrounds and normally larger sign sizes, making them easier for the model to learn and generalize. However, many critical traffic signs fall under the Regulatory and Warning categories, including speed limit, no entry, etc. As shown in Table 5, among the ten worst-performing sign types, only "Trauma Center" belongs to the Guide Signs category, with the most challenging signs coming from the Regulatory and Warning categories. This highlights significant room for improvement in the current visual model. While fine-tuned models perform well on "`Guide`" and "`Temporary Control`" signs, their performance does not consistently exceed 90%. Based on both Table 2 and Table 4 and shown in categories' accuracy, the zero-shot performance on DriveQA-V is much lower than on DriveQA-T. This indicates that current MLLMs' fine-grained perception and visual reasoning capabilities are nascent, exhibiting systematic shortcomings due to CLIP's failures.

Role of Difficulty and Distractors: To further increase the evaluation difficulty, we adopt a negative sampling strategy to construct more challenging distractors. Specifically, for DriveQA-T, we construct a difficult question set containing 1249 questions. For DriveQA-V (Signs), we leverage metadata, i.e., the ground-truth traffic sign artifact categories to ensure that distractors belong to the same category as the correct answer. For numeric signs, all candidates are constrained to numerical values to further increase ambiguity. Evaluation results on GPT-4o and a representative open-source baseline are summarized in Table 6.

Sim-to-Real Transferability: We evaluate our models finetuned on DriveQA on a curated dataset by us from Mapillary [57] (1303 annotated images, including 166 sign types), as shown in Table 7. Additionally, results in Table 8 show the downstream trajectory planning task with OpenEMMA [87] on nuScenes dataset, where our task-agnostic QA model is intentionally only fine-tuned on DriveQA but tested zero-shot in waypoint prediction to measure general-

Table 6. **Role of Difficult Questions and Distractors.** The accuracy degradation on a hard subset of DriveQA-T and on a challenging set of DriveQA-V with negative sampling shows the limitations of current models, including GPT-4o, in accurately understanding complex traffic rules and signs.

Test Set	Models	Size	Neg. Sampling		Degradation
			Before	After	
DriveQA-T	Llama-3.1 [26]	8B	55.89	39.87	28.66%
	GPT-4o [59]	-	91.96	78.91	14.19%
DriveQA-V (Signs)	LLaVA-1.5 [48]	13B	11.92	9.82	17.62%
	GPT-4o [59]	-	94.10	79.40	15.62%

Table 7. **Sim-to-Real Generalization.** We pre-train on synthetic DriveQA (DQA) and evaluate on real-world Mapillary images. The Mapillary dataset comprises challenging scenarios with various traffic sign placements, occlusion, and illumination.

Test Set	Models	Size	Accuracy	
			Off-The-Shelf	DQA-Finetuned
Real-World Mapillary [57]	Mini-InternVL [29]	2B	57.25	68.61
	LLaVA-1.5 [48]	7B	40.68	52.34
	LLaVA-1.6-mistral [49]	7B	53.18	57.71
	VILA-1.5 [47]	8B	34.38	60.86
	GPT-4o [59]	-	84.73	-

Table 8. **End-to-End Trajectory Planning Results on nuScenes.** We compute the L2 error at different prediction horizons (1s, 2s, and 3s). Lower L2 error shows our DriveQA (DQA) dataset can transfer from simulation to real-world driving tasks.

Model	Pretrained on DQA	L2(m)↓			
		1s	2s	3s	Avg.
LLaVA-1.6-mistral [49] (OpenEMMA [87])		1.49	3.38	4.09	2.98
LLaVA-1.6-mistral [49]	✓	1.30	3.46	3.98	2.91
InternVL-2.5-8B		1.66	3.36	4.15	3.06
	✓	1.30	3.08	3.73	2.71

Table 9. **Evaluation on BDD-OIA Dataset [88].** We report mean F1 score (mF1) and overall F1 score ($F1_{all}$) for both action and explanation tasks. The results show that fine-tuning on DriveQA improves performance on both tasks.

Model	Finetune		Action		Explanation	
	DQA	BDD-OIA	mF1 ↑	$F1_{all}$ ↑	mF1 ↑	$F1_{all}$ ↑
InternVL-2.5-8B			0.2951	0.554	0.0624	0.2223
	✓		0.2226	0.4103	0.1549	0.1850
		✓	0.4911	0.7072	0.2872	0.5015
	✓	✓	0.5285	0.7334	0.3102	0.5448

ization. Reduced L2 errors show the transferability of our dataset. However, nuScenes lacks diversity and is generally uneventful (e.g., minimal signage), while our benchmark exhaustively covers all traffic rules and scenarios. We therefore also make evaluations on the more diverse datasets of BDD-OIA [88] as shown in Table 9. After fine-tuning on DriveQA, the models achieve better performance in cross-domain real-world driving tasks, demonstrating the effectiveness of our data in improving the understanding of traffic rules and real-world generalizability. We provide additional analysis in the supplementary.

Limitation: While our benchmark, models, and analysis provide insights into the performance of models in understanding diverse traffic rules for autonomous driving, there are several limitations, which we plan to address in future work. First, the benchmark primarily evaluates static, structured knowledge of traffic rules. While this is aligned with standard driving knowledge tasks, there is an opportunity to leverage video-based models in the future (e.g., using our augmented CARLA simulation). Our analysis demonstrates that incorporating knowledge from text does indeed transfer to dynamic settings in nuScenes, yet vision-based reasoning remains nascent in MLLMs (or even spatial reasoning [2]). Moreover, our study highlights weaknesses in numerical reasoning and spatial awareness yet does not explore potential mitigation strategies beyond fine-tuning. The reliance on synthetic data also raises concerns about domain adaptation. Nonetheless, simulation data is crucial for scalability, as we are able to control for various variations, including occlusions and ambiguous signage, which may be rare in real-world benchmarks. Finally, while the dataset includes controlled variations in environmental factors like lighting and weather, it does not extensively cover edge cases such as emergency vehicle interactions (only covered in DriveQA-T) or pedestrian intent recognition. The models also exhibit biases towards frequently seen traffic patterns, which may result in poor generalization to geographically diverse driving environments with different road layouts and regulations.

6. Conclusion

In this paper, we introduce DriveQA, a novel benchmark for autonomous driving that evaluates models through text-based (DriveQA-T) and visual-text (DriveQA-V) question-answering, focusing on general traffic rules, traffic signs, and complex right-of-way scenarios. Our evaluation of state-of-the-art models reveals critical limitations: even fine-tuned models struggle with nuanced right-of-way scenarios, falling short of the reasoning needed for safe driving guidance. Our work deliberately focuses on static visual and textual inputs, i.e., to align with real-world driver knowledge tests. While video-based learning is not required to adhere to these standards, future research could explore hybrid frameworks incorporating video to address time-dependent scenarios. Ultimately, while humans can learn traffic rules through textual instruction and contextual practice, current models remain overly reliant on observational training data. Models thus lack the ability to internalize explicit textual knowledge and apply it effectively in decision-making. This suggests that learning traffic rules from text remains an underexplored paradigm, highlighting the need for methods that better integrate language understanding with spatial reasoning.

Acknowledgments: We thank the National Science Foundation (award IIS-2152077) and Red Hat Collaboratory (award #2024-01-RH02) for supporting this research.

References

[1] Carla autonomous driving leaderboard. *https://leaderboard.carla.org/*, 2022. 3

[2] Eyes wide shut? exploring the visual shortcomings of multimodal llms. In *CVPR*, 2024. 2, 3, 6, 8

[3] Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, et al. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv:2404.14219*, 2024. 1, 6

[4] 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:2303.08774*, 2023. 3

[5] Hidehisa Arai, Keita Miwa, Kento Sasaki, Yu Yamaguchi, Kohei Watanabe, Shunsuke Aoki, and Issei Yamamoto. Covla: Comprehensive vision-language-action dataset for autonomous driving. *arXiv:2408.10845*, 2024. 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 *CVPR*, 2020. 2, 3

[7] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, 2020. 3

[8] Ming-Fang Chang, John Lambert, Patsorn Sangkloy, Jagjeet Singh, Slawomir Bak, Andrew Hartnett, De Wang, Peter Carr, Simon Lucey, Deva Ramanan, et al. Argoverse: 3d tracking and forecasting with rich maps. In *CVPR*, 2019. 3

[9] Boyuan Chen, Zhuo Xu, Sean Kirmani, Brain Ichter, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. SpatialVLM: Endowing vision-language models with spatial reasoning capabilities. In *CVPR*, 2024. 2

[10] Feilong Chen, Minglun Han, Haozhi Zhao, Qingyang Zhang, Jing Shi, Shuang Xu, and Bo Xu. X-llm: Bootstrapping advanced large language models by treating multi-modalities as foreign languages. *arXiv:2305.04160*, 2023. 3

[11] Long Chen, Oleg Sinavski, Jan Hünermann, Alice Karnsund, Andrew James Willmott, Danny Birch, Daniel Maund, and Jamie Shotton. Driving with LLMs: Fusing object-level vector modality for explainable autonomous driving. *arXiv:2310.01957*, 2023. 1

[12] Long Chen, Oleg Sinavski, Jan Hünermann, Alice Karnsund, Andrew James Willmott, Danny Birch, Daniel Maund, and Jamie Shotton. Driving with llms: Fusing object-level vector modality for explainable autonomous driving. In *ICRA*, 2024. 3

[13] Ting Chen, Saurabh Saxena, Lala Li, David J Fleet, and Geoffrey Hinton. Pix2seq: A language modeling framework for object detection. *arXiv:2109.10852*, 2021. 3

[14] Can Cui, Yunsheng Ma, Xu Cao, Wenqian Ye, and Ziran Wang. Drive as you speak: Enabling human-like interaction with large language models in autonomous vehicles. *arXiv:2309.10228*, 2023. 1

[15] Can Cui, Yunsheng Ma, Xu Cao, Wenqian Ye, and Ziran Wang. Drive as you speak: Enabling human-like interaction with large language models in autonomous vehicles. In *WACV*, 2024. 3

[16] Can Cui, Yunsheng Ma, Xu Cao, Wenqian Ye, and Ziran Wang. Receive, reason, and react: Drive as you say, with large language models in autonomous vehicles. *IEEE ITS Magazine*, 2024. 3

[17] Can Cui, Yunsheng Ma, Xu Cao, Wenqian Ye, Yang Zhou, Kaizhao Liang, Jintai Chen, Juanwu Lu, Zichong Yang, Kuei-Da Liao, et al. A survey on multimodal large language models for autonomous driving. In *WACV*, 2024. 2

[18] 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. *arXiv:2305.06500*, 2023. 3

[19] Abhishek Das, Samyak Datta, Georgia Gkioxari, Stefan Lee, Devi Parikh, and Dhruv Batra. Embodied question answering. In *CVPR*, 2018. 2

[20] Daniel Dauner, Marcel Hallgarten, Andreas Geiger, and Kashyap Chitta. Parting with misconceptions about learning-based vehicle motion planning. *CoRL*, 2023. 3

[21] Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. CARLA: An open urban driving simulator. In *CoRL*, 2017. 2, 3, 4

[22] Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. CARLA: An open urban driving simulator. In *CoRL*, 2017. 2

[23] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 2

[24] FSD Dreams. What 'no entry' sign? https://x.com/FSDdreams/status/1781134085471048060, 2024. 2

[25] FSD Dreams. Fsd does not see or display 'no entry' signs. hhttps://x.com/FSDdreams/status/1845900114335793288, 2024. 2

[26] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv:2407.21783*, 2024. 1, 6, 8

[27] Matteo Fabbri, Guillem Brasó, Gianluca Maugeri, Orcun Cetintas, Riccardo Gasparini, Aljoša Ošep, Simone Calderara, Laura Leal-Taixé, and Rita Cucchiara. Motsynth: How can synthetic data help pedestrian detection and tracking? In *ICCV*, 2021. 3

[28] Daocheng Fu, Xin Li, Licheng Wen, Min Dou, Pinlong Cai, Botian Shi, and Yu Qiao. Drive like a human: Rethinking autonomous driving with large language models. In *WACV*, 2024. 1, 3

[29] Zhangwei Gao, Zhe Chen, Erfei Cui, Yiming Ren, Weiyun Wang, Jinguo Zhu, Hao Tian, Shenglong Ye, Junjun He, Xizhou Zhu, et al. Mini-internvl: a flexible-transfer pocket multi-modal model with 5% parameters and 90% performance. *Visual Intelligence*, 2024. 7, 8

[30] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *CVPR*, 2012. 3

[31] Maarten Grootendorst. Keybert: Minimal keyword extraction with bert., 2020. 5

[32] California Driver's Handbook. https://www.dmv.ca.gov/portal/handbook/california-driver-handbook/, 2025. 1

[33] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv:2106.09685*, 2021. 5

[34] Yihan Hu, Jiazhi Yang, Li Chen, Keyu Li, Chonghao Sima, Xizhou Zhu, Siqi Chai, Senyao Du, Tianwei Lin, Wenhai Wang, et al. Planning-oriented autonomous driving. In *CVPR*, 2023. 3

[35] Bernhard Jaeger, Kashyap Chitta, and Andreas Geiger. Hidden biases of end-to-end driving models. *arXiv:2306.07957*, 2023. 3

[36] KCEHO. Missed do not enter sign. https://x.com/KCEHO2025/status/1894844753348563046, 2025. 2

[37] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*, 2019. 5

[38] Hee Jae Kim, Kathakoli Sengupta, Masaki Kuribayashi, Hernisa Kacorri, and Eshed Ohn-Bar. Text to blind motion. *NeurIPS*, 2024. 2

[39] Lei Lai, Eshed Ohn-Bar, Sanjay Arora, and John Seon Keun Yi. Uncertainty-guided never-ending learning to drive. *CVPR*, 2024. 1

[40] Lei Lai, Zekai Yin, and Eshed Ohn-Bar. ZeroVO: Visual odometry with minimal assumptions. In *CVPR*, 2025. 3

[41] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *NeurIPS*, 2020. 4

[42] Boyi Li, Yue Wang, Jiageng Mao, Boris Ivanovic, Sushant Veer, Karen Leung, and Marco Pavone. Driving everywhere with large language model policy adaptation. In *CVPR*, 2024. 1, 3

[43] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv:2301.12597*, 2023. 3

[44] Zhiqi Li, Zhiding Yu, Shiyi Lan, Jiahan Li, Jan Kautz, Tong Lu, and Jose M Alvarez. Is ego status all you need for open-loop end-to-end autonomous driving? In *CVPR*, 2024. 2, 3

[45] Yiyi Liao, Jun Xie, and Andreas Geiger. KITTI-360: A novel dataset and benchmarks for urban scene understanding in 2d and 3d. *PAMI*, 2022. 3

[46] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, 2004. 5

[47] Ji Lin, Hongxu Yin, Wei Ping, Pavlo Molchanov, Mohammad Shoeybi, and Song Han. Vila: On pre-training for visual language models. In *CVPR*, 2024. 1, 7, 8

[48] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *CVPR*, 2024. 7, 8

[49] 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. 7, 8

[50] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *NeurIPS*, 2024. 3

[51] Yingzi Ma, Yulong Cao, Jiachen Sun, Marco Pavone, and Chaowei Xiao. Dolphins: Multimodal language model for driving. In *ECCV*, 2024. 1, 3

[52] Arjun Majumdar, Anurag Ajay, Xiaohan Zhang, Pranav Putta, Sriram Yenamandra, Mikael Henaff, Sneha Silwal, Paul Mcvay, Oleksandr Maksymets, Sergio Arnaud, et al. Openeqa: Embodied question answering in the era of foundation models. In *CVPR*, 2024. 2

[53] Jiageng Mao, Yuxi Qian, Hang Zhao, and Yue Wang. Gpt-driver: Learning to drive with gpt. *arXiv:2310.01415*, 2023. 1, 3

[54] Jiageng Mao, Junjie Ye, Yuxi Qian, Marco Pavone, and Yue Wang. A language agent for autonomous driving. In *COLM*, 2024. 1, 3

[55] Ana-Maria Marcu, Long Chen, Jan Hünermann, Alice Karnsund, Benoit Hanotte, Prajwal Chidananda, Saurabh Nair, Vijay Badrinarayanan, Alex Kendall, Jamie Shotton, et al. Lingoqa: Video question answering for autonomous driving. *arXiv:2312.14115*, 2023. 2, 3

[56] America's most trusted driver's license test prep. https://driving-tests.org/, 2025. 1, 4

[57] Gerhard Neuhold, Tobias Ollmann, Samuel Rota Bulo, and Peter Kontschieder. The mapillary vistas dataset for semantic understanding of street scenes. In *ICCV*, 2017. 2, 3, 7, 8

[58] Frank Nielsen and Frank Nielsen. Hierarchical clustering. *Introduction to HPC with MPI for Data Science*, 2016. 5

[59] OpenAI. Hello gpt-4o — openai. Retrieved in November 14, 2024 from https://openai.com/index/hello-gpt-4o/, 2024. 4, 6, 7, 8

[60] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *ACL*, 2002. 5

[61] Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. Kosmos-2: Grounding multimodal large language models to the world. *arXiv:2306.14824*, 2023. 2

[62] Renjie Pi, Jiahui Gao, Shizhe Diao, Rui Pan, Hanze Dong, Jipeng Zhang, Lewei Yao, Jianhua Han, Hang Xu, Lingpeng Kong, et al. Detgpt: Detect what you need via reasoning. *arXiv:2305.14167*, 2023. 3

[63] 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. *arXiv:2305.14836*, 2023. 2

[64] 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*, 2024. 3

[65] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 2

[66] Katrin Renz, Kashyap Chitta, Otniel-Bogdan Mercea, A. Sophia Koepke, Zeynep Akata, and Andreas Geiger. Plant: Explainable planning transformers via object-level representations. In *CoRL*, 2022. 3

[67] Katrin Renz, Long Chen, Ana-Maria Marcu, Jan Hünermann, Benoit Hanotte, Alice Karnsund, Jamie Shotton, Elahe Arani, and Oleg Sinavski. CarLLaVA: Vision language models for camera-only closed-loop driving. *arXiv preprint arXiv:2406.10165*, 2024. 3

[68] Stephan R Richter, Zeeshan Hayder, and Vladlen Koltun. Playing for benchmarks. In *ICCV*, 2017. 3

[69] 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 *WACV*, 2024. 3

[70] Hao Sha, Yao Mu, Yuxuan Jiang, Li Chen, Chenfeng Xu, Ping Luo, Shengbo Eben Li, Masayoshi Tomizuka, Wei Zhan, and Mingyu Ding. Languagempc: Large language models as decision makers for autonomous driving. *arXiv:2310.03026*, 2023. 1, 3

[71] Hao Shao, Letian Wang, Ruobing Chen, Steven L Waslander, Hongsheng Li, and Yu Liu. Reasonnet: End-to-end driving with temporal and global reasoning. In *CVPR*, 2023. 3

[72] Chonghao Sima, Katrin Renz, Kashyap Chitta, Li Chen, Hanxue Zhang, Chengen Xie, Jens Beißwenger, Ping Luo, Andreas Geiger, and Hongyang Li. Drivelm: Driving with graph visual question answering. In *ECCV*, 2024. 2, 3

[73] Yixuan Su, Tian Lan, Huayang Li, Jialu Xu, Yan Wang, and Deng Cai. Pandagpt: One model to instruction-follow them all. *arXiv:2305.16355*, 2023. 3

[74] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, et al. Scalability in perception for autonomous driving: Waymo open dataset. In *CVPR*, 2020. 3

[75] Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. Gemma 2: Improving open language models at a practical size. *arXiv:2408.00118*, 2024. 1, 6

[76] Driver Knowledge Test. https://www.driverknowledgetests.com/, 2025. 4

[77] Ran Tian, Boyi Li, Xinshuo Weng, Yuxiao Chen, Edward Schmerling, Yue Wang, Boris Ivanovic, and Marco Pavone. Tokenize the world into object-level knowledge to address long-tail events in autonomous driving. *arXiv:2407.00959*, 2024. 1

[78] 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. *CoRL*, 2024. 2, 3

[79] TT. Failed to recognize the "do not enter" sign. https://x.com/CocJii/status/1896302421862985951, 2025. 2

[80] 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:2405.01533*, 2024. 2, 3

[81] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *NeurIPS*, 2022. 4

[82] Licheng Wen, Daocheng Fu, Xin Li, Xinyu Cai, Tao Ma, Pinlong Cai, Min Dou, Botian Shi, Liang He, and Yu Qiao. Dilu: A knowledge-driven approach to autonomous driving with large language models. *arXiv:2309.16292*, 2023. 3

[83] Benjamin Wilson, William Qi, Tanmay Agarwal, John Lambert, Jagjeet Singh, Siddhesh Khandelwal, Bowen Pan, Ratnesh Kumar, Andrew Hartnett, Jhony Kaesemodel Pontes, et al. Argoverse 2: Next generation datasets for self-driving perception and forecasting. *arXiv:2301.00493*, 2023. 3

[84] Dongming Wu, Wencheng Han, Tiancai Wang, Yingfei Liu, Xiangyu Zhang, and Jianbing Shen. Language prompt for autonomous driving. *arXiv:2309.04379*, 2023. 1

[85] Yi Wu, Yuxin Wu, Georgia Gkioxari, and Yuandong Tian. Building generalizable agents with a realistic and rich 3d environment. *arXiv:1801.02209*, 2018. 2

[86] 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:2501.04003*, 2025. 2

[87] Shuo Xing, Chengyuan Qian, Yuping Wang, Hongyuan Hua, Kexin Tian, Yang Zhou, and Zhengzhong Tu. Openemma: Open-source multimodal model for end-to-end autonomous driving. In *WACV-LLVM-AD*, 2025. 2, 7, 8

[88] Yiran Xu, Xiaoyin Yang, Lihang Gong, Hsuan-Chu Lin, Tz-Ying Wu, Yunsheng Li, and Nuno Vasconcelos. Explainable object-induced action decision for autonomous vehicles. In *CVPR*, 2020. 2, 8

[89] 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. *RA-L*, 2024. 1, 3

[90] Lu Yuan, Dongdong Chen, Yi-Ling Chen, Noel Codella, Xiyang Dai, Jianfeng Gao, Houdong Hu, Xuedong Huang, Boxin Li, Chunyuan Li, et al. Florence: A new foundation model for computer vision. *arXiv:2111.11432*, 2021. 3

[91] Xiaohua Zhai, Xiao Wang, Basil Mustafa, Andreas Steiner, Daniel Keysers, Alexander Kolesnikov, and Lucas Beyer. Lit: Zero-shot transfer with locked-image text tuning. In *CVPR*, 2022.

[92] Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv:2306.02858*, 2023. 3

[93] Jimuyang Zhang, Ruizhao Zhu, and Eshed Ohn-Bar. SelfD: self-learning large-scale driving policies from the web. In *CVPR*, 2022. 3

[94] Jimuyang Zhang, Zanming Huang, and Eshed Ohn-Bar. Coaching a teachable student. In *CVPR*, 2023. 3

[95] Jimuyang Zhang, Zanming Huang, Arijit Ray, and Eshed Ohn-Bar. Feedback-guided autonomous driving. In *CVPR*, 2024. 1, 2

[96] Zhejun Zhang, Alexander Liniger, Dengxin Dai, Fisher Yu, and Luc Van Gool. End-to-end urban driving by imitating a reinforcement learning coach. In *ICCV*, 2021. 3

[97] Baichuan Zhou, Ying Hu, Xi Weng, Junlong Jia, Jie Luo, Xien Liu, Ji Wu, and Lei Huang. Tinyllava: A framework of small-scale large multimodal models. *arXiv:2402.14289*, 2024. 1, 2

[98] Yunsong Zhou, Linyan Huang, Qingwen Bu, Jia Zeng, Tianyu Li, Hang Qiu, Hongzi Zhu, Minyi Guo, Yu Qiao, and Hongyang Li. Embodied understanding of driving scenarios. *ECCV*, 2024. 1, 3, 4

[99] Ruizhao Zhu, Peng Huang, Eshed Ohn-Bar, and Venkatesh Saligrama. Learning to drive anywhere. *CoRL*, 2023. 3