Supplementary Material for
Passing the Driving Knowledge Test

Maolin Wei*, Wanzhou Liu*, Eshed Ohn-Bar

Abstract

*Our main paper introduced the first comprehensive benchmark for reasoning over traffic rules and regulations. In this supplementary material, we provide additional details on two key aspects of DriveQA. First, we elaborate on the **dataset construction process**. Second, we present further analysis, including **failure cases and dataset ablations, as well as evaluations on real-world downstream tasks**.*

Contents

1. DriveQA Dataset and Implementation

1.1. Data Collection

DriveQA-T Dataset: To generate a diverse set of textual questions, we begin by manually writing a small set of 1,000 diverse questions. We then use these as prompt to GPT-4o Fig. S1 (through the API) to sample a larger dataset of questions. Out of this large set, we filter out redundant samples and incorrect questions resulting in 25K clean QA pairs with correct explanations. We further determine an initial type for each question through hierarchical clustering over BERT [8] embeddings (with 19 clusters) and subsequently verify the assignments for samples that are far from the centroids.

DriveQA-V Dataset: Our image-and-text DriveQA-V dataset comprises two primary types of scenarios: *traffic signs* and *right-of-way*. The dataset images were collected using the open-source CARLA simulator [4], for which we developed custom scripts to procedurally generate VQA pairs. We introduced two key modifications to CARLA to achieve this.

Based on the U.S. state and corresponding driver's manual context, please generate different multi-choice questions for the driver permit examination with correct answers and brief explanations. These questions should cover different categories: intersection, lane change, lights, parking, regulation, signal, signs, speed, symbols, traffic, alcohol or drugs, and right of way.

Please follow the format of the example question:

Question: What should you do if you plan to pass another vehicle?
A. Assume the other driver will let you pass as long as you signal
B. Assume there is nothing at your blind spot without doing a shoulder check
C. Do not assume the other driver will make space for you to pass
D. Assume there is no car behind you

Correct Answer: C. Do not assume the other driver will make space for you to pass

Explanation: You should not assume other vehicles will make space for you to pass. Never overtake and pass another vehicle unless you are sure you can do so without danger to yourself or others.

Figure S1. **Prompt Example Data Generation.** GPT-4o prompt for generating the preliminary DriveQA-T data. The data is then inspected and filtered manually to ensure high quality.

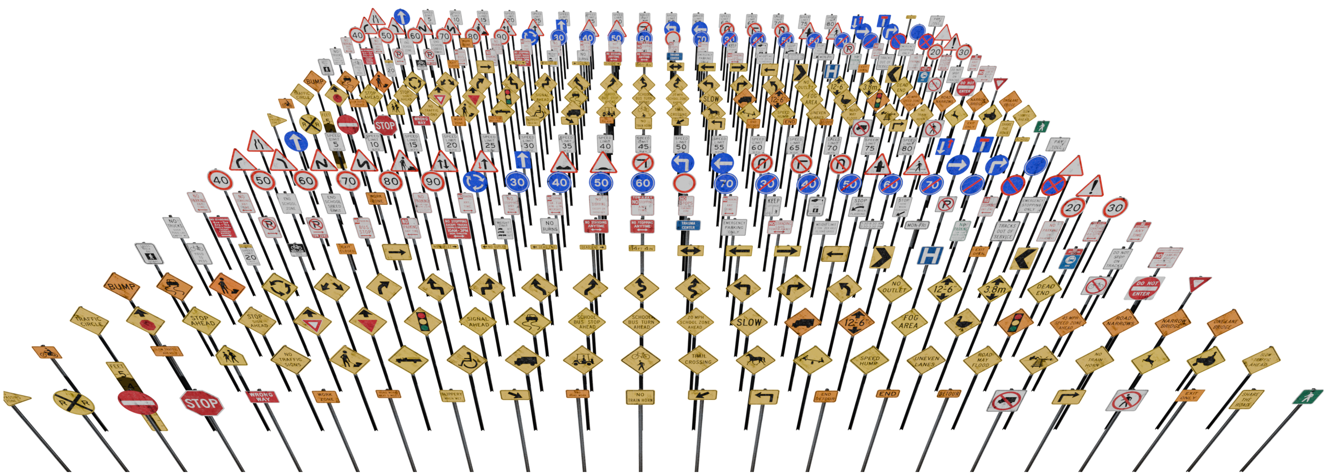


Figure S2. **Traffic Sign Assets Imported Into CARLA.** Originally, CARLA has two types of signs (yield and stop sign). We augment the simulation by importing a diverse set of 3D traffic sign assets.

Traffic Signs: First, CARLA's default traffic sign library and metadata are lacking as they contain only two sign types (yield and stop signs). To address this, we incorporated 220 3D models of U.S. ***traffic signs*** into CARLA's blueprint library. Figure S2 illustrates all the traffic sign prototypes added to our study. These signs were strategically distributed along both sides of the roads in the simulator, and a vehicle equipped with a forward-facing camera was used to capture images of the signs. This method enabled us to collect diverse traffic sign images with corresponding attribute metadata, including weather conditions, capture distances, camera heights, and town locations. The dataset comprises 220 unique traffic sign designs, with each sign available in two variations: a clean version and a stained version, resulting in a total of 440 options. Based on the classification standards commonly outlined in the driver's manuals of most U.S. states, we categorized the signs into 4 types: Regulatory Signs, Warning Signs, Guide Signs, and Temporary Traffic Control Signs. These categories are visually distinguishable by their color schemes: Temporary Traffic Control Signs feature an orange background, Guide Signs use blue and green backgrounds, Warning Signs are characterized by a yellow background, and all remaining signs fall under Regulatory Signs. After classification, the dataset includes 102 Regulatory Signs, 70 Warning Signs, 26 Guide Signs, and 22 Temporary Traffic Control Signs. We assign into types by intersection type, camera perspective, and sign type (e.g., "*Regulatory*", "*Warning*", "*Guide*" and "*Temporary Control*"). When building QA pairs, we leverage ground truth sign category information to make our data better aligned with real driving exam. Specifically, all numeric QAs related with speed limitation are combined with one right speed choice and three speed distractors. Besides numeric traffic sign QA pairs, other

QA pairs are also constructed with the distractors in the same sign category with the right answer, making it hard to choose the right answer. More illustrations are shown in Fig. 2 in the main paper.

Right-of-Way: Second, to simulate ***right-of-way*** traffic scenarios at intersections, we developed an automated spawning and classification algorithm leveraging the simulator's ground-truth state. We capture and save images from both a top-down view and the ego-vehicle perspective. The questions involve determining right-of-way rules, with tasks ranging from simple right-of-way identification to more complex sequences that describe the order of vehicle movement. For each intersection scenario, we adjust the number of vehicles and assign directions for each vehicle. Specifically, CARLA's map topology was utilized to identify pre-turn waypoints by backtracking from junctions within a configurable distance range. Vehicles were then spawned at these waypoints to ensure their positions were accurately set before entering the intersection. We randomly spawned 1–4 vehicles per intersection and assigned their driving directions (e.g., straight, left, or right) based on the angular differences between their current and subsequent waypoints. Each vehicle was assigned a unique color for clarity. We collect images both from first-person and top-down views. To construct right-of-way QA pairs, we implemented a script to automatically generate two question types: "`Who has the right-of-way at the intersection?`" and "`In which order should they proceed?`". These questions were created based on all possible combinations of vehicles and their assigned driving directions at the intersection, allowing for the generation of up to dozens of questions per intersection. Answers were determined based on standard driver handbook rules, specifically:

- Rule 1 - The vehicle that arrived first has the right-of-way.
- Rule 2 - If two or more vehicles arrive at an intersection simultaneously, drivers on the left must yield to drivers on the right.
- Rule 3 - A vehicle turning left must yield to oncoming traffic, even if it arrives first.
- Rule 4 - If two vehicles are both turning left, they may turn without yielding by passing in front of each other.

Dataset Split: For the DriveQA-T dataset, we split the data into 18K QA pairs for training, 4K for validation, and 4K for testing, ensuring that each subset contains a representative distribution of all question types. For the DriveQA-V dataset, we sample 40K intersection QA pairs for training and 8K for testing, as well as 20K sign QA pairs for training and 5K for testing. To evaluate the model's performance in varying background settings, we ensure that the testing data includes samples from several distinct towns in CARLA. For the DriveQA-T dataset, we applied a 7:1.5:1.5 split into training, validation, and test sets based on the types of questions, ensuring each set covers all types of questions. For the DriveQA-V dataset, we adopt distinct strategies for traffic signs and right-of-way scenarios. In the traffic sign data, we use 85% of the data from Town03, Town05, and Town07 for training and validation, within 15% of this subset as the validation set. We retain 15% of the data from these towns as an in-scene test set to assess generalization within the same environment. For a broader evaluation, we designate all data from Town01 and Town10 as an independent test set, providing a diverse range of conditions for comprehensive model assessment. The right-of-way scenarios required tailored divisions due to the variety of intersection types and quantities across different CARLA maps. For cross intersections, we select Town04 and Town05, which contain a larger number of cross intersections as the training set. Town03 is used for validation, while Town07 and Town10 serve as the test set. For V(Inters.), we apply a similar strategy. We select Town01, Town02, and Town07 as the training set, and use Town04 as the validation set, while Town10 and Town05 are served as the test set. This well-structured dataset division ensures that models are trained on a representative variety of scenarios while reserving unique data for validation and testing, enabling a thorough assessment of model generalization across unseen traffic environments.

Table S1. **Challenging Categories on DriveQA-T**. We show the results of most difficult 10 types: Limits: Speed and Distance Limits, Alcohol: Blood Alcohol Limits and DUI Laws, Passing: Passing Rules and Lane Usage in Restricted Situations, Penalties: Driver's License Penalties, Parking: Parking and Wheel Positioning, Highway: Passing Rules and Lane Usage in Highway, Turning: Turning Rules, Signs: Traffic Signs and Signals, Headlight: Headlight Usage, Intersection: Right-of-Way and Lane Selection. We denote with **green the top method**, and **light green second best**.

Models	Size	CoT	RAG	Finetune	Limits	Alcohol	Passing	Penalties	Parking	Highway	Turning	Signs	Headlight	Intersection	Average
Gemma-2 [16]	2B				42.15	62.96	37.28	46.89	35.64	45.71	48.39	42.54	52.02	27.88	44.15
		✓			42.98	72.22	53.05	55.60	42.57	54.29	54.84	56.91	60.69	34.51	52.77
		✓	✓		58.68	79.63	45.16	59.75	47.52	47.62	52.69	55.80	63.58	55.75	56.62
		✓	✓	✓	62.40	70.37	69.18	72.61	61.39	74.76	70.43	76.24	76.88	85.84	72.01
Gemma-2 [16]	9B				57.85	74.07	71.33	79.25	54.46	75.71	72.58	77.90	88.44	58.41	71.00
		✓			59.50	81.48	69.18	79.67	58.42	73.81	73.66	78.45	84.97	62.83	72.20
		✓	✓		64.88	83.33	77.42	73.44	68.32	76.19	74.73	86.74	86.13	77.88	76.91
		✓	✓	✓	72.31	83.33	88.53	86.72	88.12	88.57	87.63	92.82	93.64	91.15	87.28
Llama-3.1 [5]	8B				53.72	77.78	55.56	58.51	37.62	48.10	57.53	53.04	68.79	48.23	55.89
		✓			55.37	75.93	53.05	53.11	38.61	47.14	48.39	54.70	69.94	65.93	56.22
		✓	✓		55.37	72.22	53.76	61.83	46.53	49.52	54.84	72.38	72.83	68.58	60.79
		✓	✓	✓	72.73	88.89	89.25	88.80	86.14	84.76	90.86	89.50	93.64	91.59	87.62
Llama-3.2 [5]	3B				36.78	57.41	54.84	63.90	35.64	50.00	50.54	51.93	65.32	42.92	50.93
		✓			48.35	74.07	46.59	53.94	26.73	43.81	43.55	43.65	58.96	49.56	48.92
		✓	✓		61.16	72.22	65.95	66.80	53.47	62.38	55.91	68.51	73.99	61.50	64.19
		✓	✓	✓	69.42	81.48	82.08	83.40	75.25	85.71	86.02	91.16	87.86	85.84	82.82
Phi-3.5-mini [1]	3.8B				49.17	72.22	69.53	78.01	48.51	73.81	73.66	70.72	82.66	79.65	69.79
		✓			55.79	75.93	69.89	79.67	45.54	78.57	76.88	67.40	82.08	79.65	71.14
		✓	✓		63.22	74.07	75.27	82.57	62.38	78.57	80.65	83.98	87.28	84.96	77.30
		✓	✓	✓	66.94	81.48	81.00	83.82	65.35	86.19	86.56	87.85	84.39	87.17	81.08
GPT-4o [14]	-	✓	✓		76.72	88.89	92.59	90.76	93.75	93.27	92.05	94.32	100.00	97.27	91.96

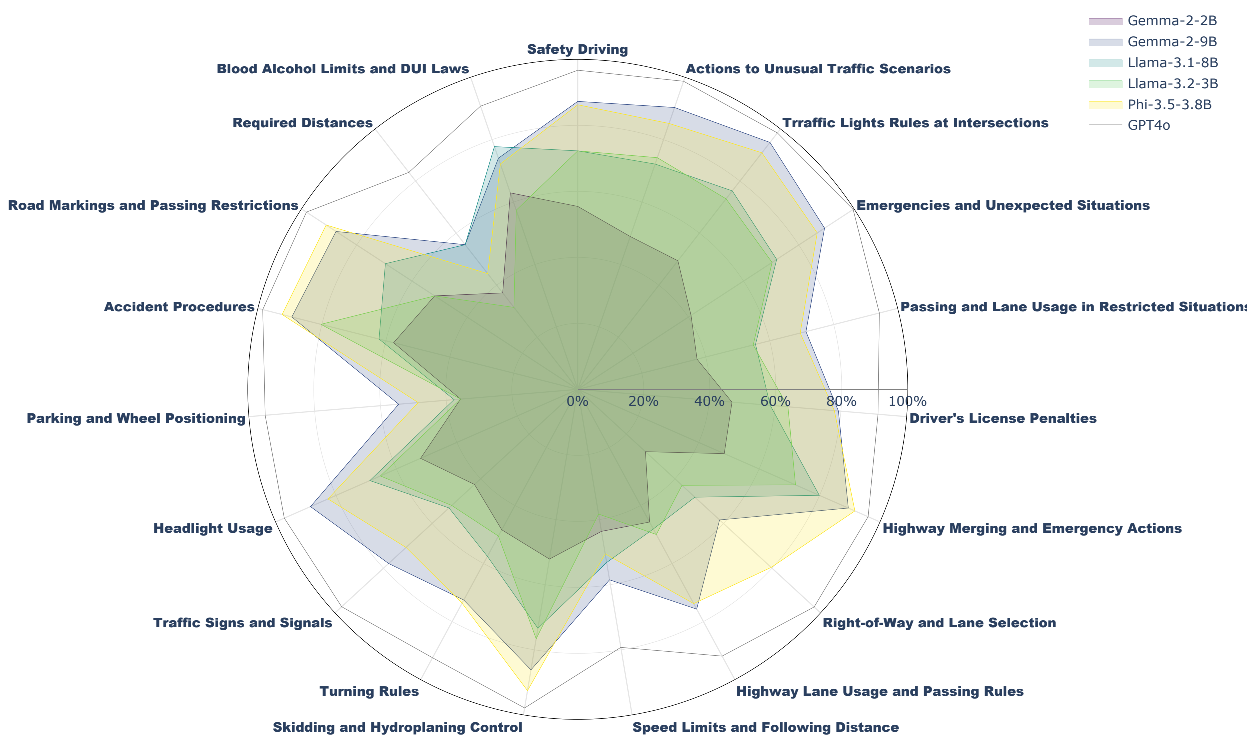


Figure S3. **DriveQA-T Accuracy Breakdown.** We analyze model performance for different categories of questions.

Prompt to the Model without CoT:
Given a multiple-choice question commonly found on a US driver's license test, select the correct answer option and provide only the correct answer option. Do not provide any explanation for the answer.

Context: {RAG_context} (Adding context if using RAG)

Prompt to the Model with CoT:
Given a multiple-choice question commonly found on a US driver's license test, think step by step, then select the correct answer option and provide your reasoning for the answer. Do not mention options except the correct option in the response.

Context: {RAG_context} (Adding context if using RAG)

Figure S4. **Prompt Example for Model Input.** We show the prompt formats used with and without chain-of-thought reasoning and optional RAG-based context retrieval.

1.2. Dataset Statistics and Characteristics

Our final DriveQA-T dataset consists of 26,143 text-based question-answer pairs derived from 51 official U.S. driver's handbooks (50 states plus Washington, D.C.), organized into five major categories and 19 sub-classes. Each pair includes detailed explanations for the correct answers. DriveQA-V contains 68K images and 448K QA pairs generated using the CARLA simulator across seven distinct maps featuring diverse weather and lighting conditions. The DriveQA-V (Signs) includes 48K CARLA traffic sign images, by leveraging 220 U.S. traffic sign 3D models, and 1,303 real world traffic sign images in Mapillary Dataset, all paired with questions and answers. The DriveQA-V (Interns.) contains 20K intersection images accompanied by 400K right-of-way questions and answers. These questions cover both T-intersections and cross-intersections, captured from multiple camera perspectives, including top-down and first-person views. The visual dataset questions are systematically categorized by sign types (Regulatory, Warning, Guide, and Temporary Control) and intersection scenarios (based on intersection type and camera view). This structure makes the dataset particularly well-suited for evaluating both textual understanding of traffic rules and visual reasoning in complex traffic scenarios.

1.3. Model Training Protocol

All of our fine-tuning is implemented with LoRA [7]. We adopt a learning rate of 2e-5, a LoRA alpha of 32, and a LoRA rank of 16 for Text LLMs. For MLLMs such as LLaVA-1.5 [11] and InternVL-2.5 [3], we use the official default configurations. Experiments are conducted with 1-3 L40S GPUs. Regarding efficiency, training one epoch on DriveQA with InternVL-2.5 and CoT+RAG takes 13 hours. Inference takes 2s/image with CoT and 0.5s without (0.3s for LLaMA-Adapter). Additionally, the input prompt to the model is show in Fig. S4.

2. Additional Experiments

2.1. Scenario Type Breakdown on DriveQA-T

In the main paper, we analyze model performance across a subset of the different types of questions. The 10 most difficult types are illustrated in Table S1. Our radar plot in Fig. S3 shows the complete analysis over all 19 types of questions in the dataset. We find models performance to vary significantly across question types, i.e., questions related to distances, positioning, traffic rules, speed limits are particularly challenging overall.

2.2. Traffic Sign Performance Breakdown by Type

Based on the accuracy distribution shown in Table S2, we observe a distinct pattern in the model's performance across various types of traffic signs. A significant majority of signs (47.73%) achieve high accuracy rates above 90%, with 28 signs recognized perfectly (100%) and 77 signs falling within the 90–99% range. These high-performing signs generally exhibit characteristics such as simple geometric shapes, high-contrast colors, and frequent appearances in the training data, as seen in signs like "Straight Ahead" and "Stop." The model demonstrates moderate performance (80–89%) on signs with more intricate elements, such as those displaying speed limits or directional information, which constitute 30.91% of all signs. Notably, accuracy decreases significantly for signs containing multiple visual components or text-heavy content. Only a small fraction of signs (9.09%) fall below 70% accuracy, primarily those involving complex scenarios or rare occurrences. The lowest-performing signs, "including Tractor Crossing" and "Trauma Center" (with accuracies below 60%), are characterized

Table S2. **Traffic Sign Performance by Type.** We categorize traffic signs by accuracy range and report representative characteristics. For each range, we show the number of signs, representative examples, and key observations about sign characteristics in that range. Key characteristics describe common features of signs within each accuracy range that may influence model performance. We note that the accuracy ranges are based on the performance of our best-performing model (VILA-1.5 after fine-tuning).

Accuracy Range	# Signs	% of Total	Representative Examples	Key Characteristics
100.00%	28	12.73%	Straight Ahead Traffic Light Ahead Exit Only	Simple geometric shapes High contrast colors Common in training data
90-99%	77	35.00%	Stop No Parking No U-Turns	Simple geometric shapes High contrast colors Common in training data
80-89%	68	30.91%	Double Bend, First to Right 20 MPH School Zone Ahead Speed Limit	Contains directional information Multiple text elements Standard color schemes
70-79%	27	12.27%	Trail Crossing Stop Here When Flashing Handicap Bus Stop, No Standing	More complex symbols Composite visual elements Mixed text and symbols
60-69%	18	8.18%	No Parking Any Time Bicycles Keep Left, Pedestrians Keep Right Low Ground Clearance	Text-heavy signs Complex scenarios Multiple visual components
<60%	2	0.91%	Tractor Crossing Trauma Center -	Complex symbols Rare or unusual signs -

Table S3. **Role of Environmental Conditions.** We report overall accuracy breakdown over different times of day and cities in CARLA. We find performance over times of day is shown to be consistent, yet some cities are more challenging than others (Town03, Town10).

Condition	Dawn	Morning	Night	Noon	Sunset	Twilight	Avg.
Town01	93.85	91.57	92.11	93.94	90.84	93.49	92.63
Town03	73.59	73.52	74.22	70.64	75.75	79.73	74.58
Town05	84.52	81.02	80.07	81.99	78.32	78.91	80.81
Town07	91.18	91.32	93.47	91.88	88.34	91.82	91.34
Town10HD	66.29	60.24	66.11	66.11	66.38	66.17	65.22
Avg.	81.89	79.53	81.20	80.91	79.93	82.02	80.91

by their unique symbols and infrequent presence in real-world settings. This distribution indicates that while the model excels at recognizing common and geometrically simple signs, there remains substantial room for improvement in handling traffic signs with higher complexity or lower prevalence.

2.3. Weather, Daylight, Town, and Capture Distance Settings

We analyze across different conditions in Table S3, Table S4, and Table S5. We note that our dataset is split by reserving Town01 and Town10 for testing while training on Town03, Town05, and Town07, which represent small-town, urban, and rural environments. We observe sensitivity to town complexity, e.g., Town10, the most complex and realistically rendered town, exhibits lower performance. Furthermore, while the model demonstrates variable performance across daylight conditions, e.g., worse performance in high illumination and reflection conditions. These findings surface challenges in existing MLLMs to guide future research.

2.4. Roundabout Evaluation

We collect a small set of 100 images for roundabout and create 500 questions for an additional evaluation, as a complementary setting to the intersection questions in our DQA-V. Table S6 shows that the models fine-tuned on DQA can also generalize

Table S4. **Weather and Times of Day Combinations.** We report model accuracy over varying weather and times of day conditions. We find after rain reflections (wet conditions) to be more challenging than other weathers. Morning and sunset present challenging illumination conditions. Results are shown using the VILA-1.5-8B.

Condition	**Dawn**	**Morning**	**Night**	**Noon**	**Sunset**	**Twilight**	**Avg.**
Clear	87.69	85.19	88.80	88.96	89.69	89.65	88.33
Cloudy	89.26	82.34	88.93	85.92	87.24	89.43	87.19
HardRain	91.26	87.84	88.70	90.00	77.89	88.09	87.30
MidRain	88.15	85.62	91.82	87.10	89.29	88.96	88.49
SoftRain	93.27	91.44	86.14	92.07	88.14	87.72	89.80
Wet	81.64	78.28	87.70	82.23	84.96	85.77	83.43
WetCloudy	88.04	84.00	84.84	88.19	85.16	86.74	86.16
Avg.	88.47	84.96	88.13	87.78	86.05	88.05	87.24

Table S5. **Impact of Daylight Conditions on Right-of-Way VQA Pairs.** Results are reported using finetuned LLaVA and VILA.

Model	Dawn	Morning	Night	Noon	Sunset	Twilight
T-Intersection With Vehicle's Perspective						
LLaVA-1.5-7B [11]	62.38	66.13	63.31	65.24	66.67	61.50
VILA-1.5-8B [10]	49.50	44.09	47.93	50.27	46.99	47.06
GPT-4o [14]	63.89	66.67	52.17	59.38	50.00	35.71
T-Intersection With Top-Down Perspective						
LLaVA-1.5-7B [11]	71.81	63.55	65.60	69.02	84.04	68.94
VILA-1.5-8B [10]	54.81	56.04	53.08	54.90	44.12	50.82
GPT-4o [14]	55.22	63.16	60.29	60.56	52.78	72.13
Cross-Intersection With Vehicle's Perspective						
LLaVA-1.5-7B [11]	56.98	49.52	58.07	57.25	51.95	53.67
VILA-1.5-8B [10]	54.34	57.88	57.22	57.25	56.25	50.44
GPT-4o [14]	50.00	49.02	47.27	50.88	56.76	51.16
Cross-Intersection With Top-Down Perspective						
LLaVA-1.5-7B [11]	62.50	55.21	52.78	56.95	57.58	54.19
VILA-1.5-8B [10]	59.14	58.33	59.26	56.95	52.46	57.26
GPT-4o [14]	70.51	60.78	58.44	57.32	48.00	59.80

Table S6. **Performance of MLLMs on Roundabout Questions.** We construct a set of roundabout questions as a complement to the intersection questions.

Model	**Finetuned on DQA**	**Accuracy**
LLaVA-1.5 [11]		45.33
	✓	**77.78**
LLaVA-1.6-mistral [12]		54.44
	✓	**78.89**

to roundabout questions.

2.5. Real-World Impact and Cross-Dataset Analysis

Furthermore, we quantify the impact of pre-training on DriveQA for downstream tasks by evaluating on *DriveLM-nuScenes* decision-making QA as shown in Table S7 and Table S8, and on other diverse datasets, BDD-X [9] and MAPLM [2], as shown in Table S9 and Table S10. These results collectively validate that DriveQA effectively bridges the sim-to-real gap,

Table S7. **DriveQA Pre-Training Improves Downstream Performance.** We evaluate the impact of DriveQA pre-training and evaluate on DriveLM [15]. The results demonstrate that DriveQA incorporates generalized knowledge applicable to various downstream driving tasks, e.g., for perception and planning.

Models	**Size**	**Finetuning**			**Accuracy**
		DriveLM	**DriveQA-T**	**DriveQA-V**	
LLaMA-Adapter-V2 [6]	7B	✓			60.21
		✓	✓		61.13
		✓		✓	61.80
		✓	✓	✓	**63.27**

Table S8. **DriveQA vs. DriveLM.** We train with the DriveLM, a related QA benchmark, on DriveQA to highlight the difference between the two QA tasks. We observe off-the-shelf model generalization of LLaMA-Adapter-v2 to DriveQA is poor. While fine-tuning the model improves performance for our right-of-way and sign understanding tasks, pre-training on DriveLM does not benefit our task.

Models	**Size**	**Pretraining on DriveLM**	**fine-tuning on DriveQA-V**	**Intersection**	**Sign**
LLaMA-Adapter-v2 [6]	8B			22.05	25.85
			✓	75.44	31.71
		✓	✓	68.67	31.65

Table S9. **Evaluation on BDD-X Dataset [9].** We show high-level action (BLEU4, CIDEr) and low-level steering angle (in degrees) and speed (in m/s) prediction.

Model	**Finetune**		**High-Level Action**		**Speed**	**Steer**
	DQA	BDD-X	BLEU4 ↑	CIDEr ↑	RMSE ↓	RMSE ↓
InternVL-2.5-8B [3]			3.7	11.0	1.47	5.75
	✓		3.0	11.0	1.62	5.44
		✓	30.8	214.3	0.66	4.03
	✓	✓	**32.1**	**224.1**	**0.64**	**4.00**

Table S10. **Evaluation on MAPLM Dataset [2].** We report performance across five question types: SCN (road scene), QLT (data quality), LAN (lane number), INT (road cross), and DES (lane attribute description). We follow [2] and use metrics including accuracy for each type, along with overall frame-level accuracy (FRM) and question-level accuracy (QNS). We find that DriveQA can significantly improve the accuracy of LAN, and fine-tuning on DriveQA and MAPLM achieves the best overall results.

Model	**Finetune**		**Open QA**		**Fine-grained QA**			**FRM ↑**	**QNS ↑**
	DQA	MAPLM	LAN ↑	DES ↑	INT ↑	QLT ↑	SCN ↑		
InternVL-2.5-8B			49.93	0.00	76.67	12.73	86.80	4.40	56.53
	✓		60.40	0.00	76.53	12.40	82.33	4.07	57.92
		✓	**100.00**	**75.43**	**78.40**	83.40	96.13	63.73	89.48
	✓	✓	**100.00**	74.25	78.20	**83.60**	**96.53**	**64.00**	**89.58**

overcoming fundamental limitations of purely synthetic training paradigms while enhancing real-world generalization.

Additionally, we report an ablation leveraging the LLaMA-Adapter-v2 model and DriveLM dataset [15] (Table **??**). We find that pre-training on the DriveLM dataset does not transfer knowledge to the DriveQA-V task, i.e., in contrast to direct fine-tuning. The worse performance suggests that having access to more data does not necessarily lead to better performance.

2.6. Additional Qualitative Examples

We depict additional examples from our DriveQA dataset in Fig. S6 for pure text questions, Fig. S7, Fig. S8, Fig. S9, Fig. S10, Fig. S11 for intersection questions, Fig. S12, Fig. S13, Fig. S14 for signs understanding questions, and Fig. S15, Fig. S16 for Mapillary extended evaluation.

2.7. Additional Quantitative Results

We place the fine-tuning results of LLaVA-1.5 [11], on all 220 traffic signs in Table S11, Table S12, and Table S13.

Figure S5. **Example Real Sign Images of Mapillary Dataset.** We construct an extended traffic scene dataset with the similar complexity and ensure each frame containing one valid sign by filtering Mapillary [13]. More failure cases are illustrated in Fig. S15 and Fig. S16.

Which of the following statements is true about making a left turn at a traffic light with a green left arrow?

A. You must yield to oncoming traffic. B. You have the right-of-way and can make the turn without stopping.
C. You must come to a complete stop before turning. D. None of the above.

Explanation: A green left arrow indicates that oncoming traffic is stopped by a red light, allowing you to turn left without yielding.

While merging onto an expressway, what is essential for maintaining safety and smooth traffic flow?

A. Stopping at the end of the acceleration lane until an opening appears. B. Accelerating quickly to any speed as long as you are ahead of surrounding traffic. C. Matching your speed with highway traffic and merging smoothly.
D. Honking loudly to alert highway drivers you're merging.

Explanation: Merging at similar speeds as existing traffic helps maintain the flow and prevents potential accidents due to sudden speed differences.

In which of the following situations is it legal to pass another vehicle?

A. On a two-lane road with a solid yellow line on your side. B. At intersections.
C. When approaching the crest of a hill. D. None of the above.

Explanation: Passing is not permitted in these situations due to visibility issues and potential hazards that may not be immediately apparent.

When is it permissible to pass another vehicle on the right?

A. When the vehicle ahead is turning left. B. On a one-way street with two or more lanes.
C. On a limited-access highway with two or more lanes in one direction. D. All of the above.

Explanation: Passing on the right is allowed under certain conditions such as when a vehicle is making a left turn, on one-way streets with multiple lanes, or on highways with multiple lanes in one direction.

The minimum speed on a New Hampshire interstate highway during safe driving conditions is:

A. 45 MPH unless otherwise posted. B. 35 MPH on all occasions.
C. Determined solely by weather conditions. D. No minimum speed limit exists.

Explanation: The state law designates a minimum speed limit of 45 MPH during safe driving conditions to maintain traffic flow unless certain conditions dictate otherwise.

Figure S6. **More Example Questions and Answers of DriveQA-T Dataset.** We random sample some failure cases of the text language model Llama-3.1 [5] based on our evaluation of DriveQA-T. Each text question in DriveQA-T has the attribute: state name from the reference driver's handbook, the question type, question, choices, correct answer and explanation.

Question: Assuming several cars arrive at an uncontrolled intersection simultaneously. The Red car will go straight. The Purple car will turn left. The Blue car will turn left. **In which order should they proceed?**

A. Red car -> Purple car -> Blue car
B. Red car -> Blue car -> Purple car
C. Purple car -> Blue car -> Red car
D. Blue car -> Purple car -> Red car

LLaVA-1.5
D. Blue car -> Purple car -> Red car

LLaVA-1.6
D. Blue car -> Purple car -> Red car

VILA-1.5
A. Red car -> Purple car -> Blue car

Mini-InternVL
B. Red car -> Blue car -> Purple car

GPT-4o
A. Red car -> Purple car -> Blue car

Question: Assuming several cars arrive at an uncontrolled intersection simultaneously. The Yellow car will go straight. The Blue car will turn left. **In which order should they proceed?**

A. Yellow car -> Blue car
B. Blue car -> Yellow car
C. No one needs to yield because there is no conflict
D. None of the above

LLaVA-1.5
B. Blue car -> Yellow car

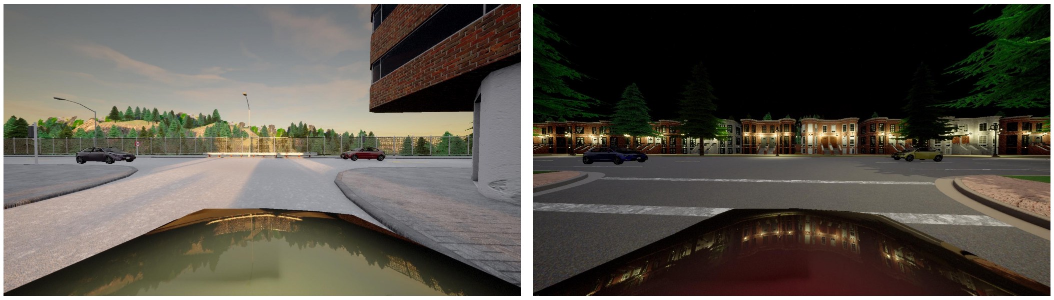
LLaVA-1.6
C. No one needs to yield because there is no conflict

VILA-1.5
C. No one needs to yield because there is no conflict

Mini-InternVL
A. Yellow car -> Blue car

GPT-4o
A. Yellow car -> Blue car

Figure S7. **T-Intersection Q&A Examples from Top-down Perspective and Model Predictions.** In green is the ground truth answers. In red is the error made by the model.



Question: Assuming several cars arrive at an uncontrolled intersection simultaneously. The Purple car will go straight. The Red car will turn left. The ego car will turn left. **In which order should they proceed?**

A. Purple car -> Red car -> Ego car
B. Red car -> Purple car -> Ego car
C. Purple car -> Ego car -> Red car
D. Ego car -> Red car -> Purple car

LLaVA-1.5
C. Purple car -> Ego car -> Red car

LLaVA-1.6
C. Purple car -> Ego car -> Red car

VILA-1.5
A. Purple car -> Red car -> Ego car

Mini-InternVL
B. Red car -> Purple car -> Ego car

GPT-4o
A. Purple car -> Red car -> Ego car

Question: Assuming several cars arrive at an uncontrolled intersection simultaneously. The Blue car will turn right. The ego car will turn right. The Yellow car will go straight. **Who has the right-of-way at the intersection?**

A. Blue car
B. Ego car
C. Yellow car
D. No one needs to yield because there is no conflict.

LLaVA-1.5
B. Ego car

LLaVA-1.6
A. Blue car

VILA-1.5
A. Blue car

Mini-InternVL
D. No one needs to yield because there is no conflict.

GPT-4o
D. No one needs to yield because there is no conflict.

Figure S8. **T-Intersection Q&A Examples from Vehicle's Perspective Perspective and Model Predictions.** In green is the ground truth answers. In red is the error made by the model.

Question: Assuming in an uncontrolled intersection, the Yellow car arrives later, other cars arrive earlier and simultaneously. The Purple car will go straight. The Blue car will turn left. The Yellow car will turn right. The Red car will go straight. **Who has the right-of-way at the intersection?**

A. Purple car
B. Blue car
C. Red car
D. Yellow car

LLaVA
A. Purple car

VILA
D. Yellow car

GPT-4o
A. Purple car

Question: Assuming several cars arrive at an uncontrolled intersection simultaneously. The Red car will turn left. The Blue car will turn right. **Who has the right-of-way at the intersection?**

A. Red car
B. Blue car
C. No one needs to yield because there is no conflict
D. None of the above

LLaVA
C. No one needs to yield because there is no conflict

VILA
C. No one needs to yield because there is no conflict

GPT-4o
B. Blue car

Figure S9. **Cross-Intersection Q&A Examples from Top-down Perspective and Model Predictions.** In green is the ground truth answers. In red is the error made by the model.

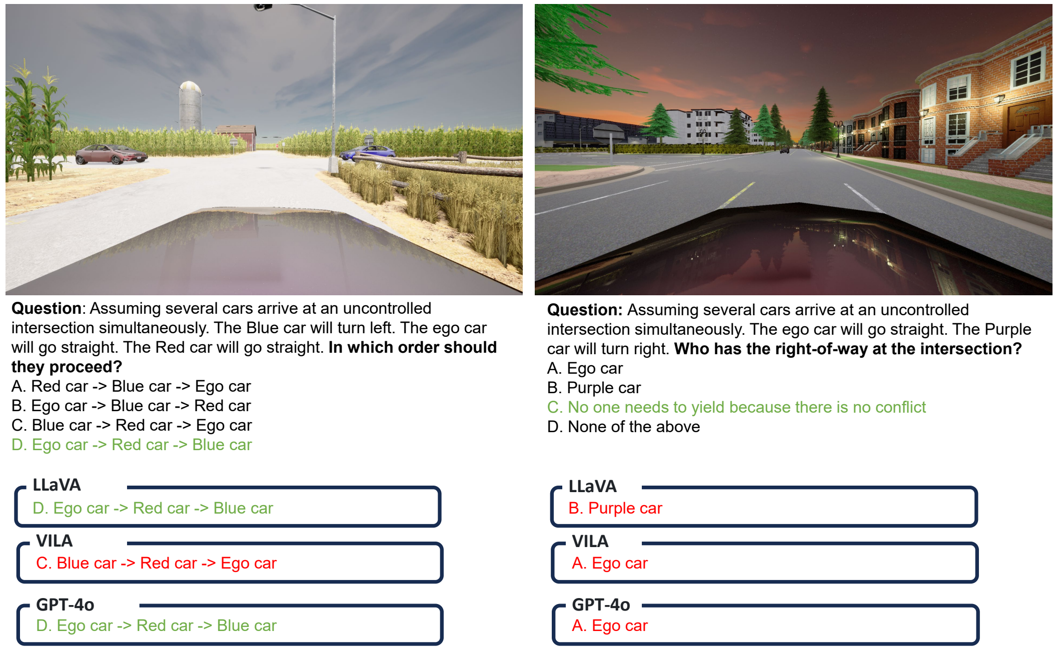


Figure S10. **Cross-Intersection Q&A Examples from Vehicle's Perspective and Model Predictions.** In green is the ground truth answers. In red is the error made by the model.

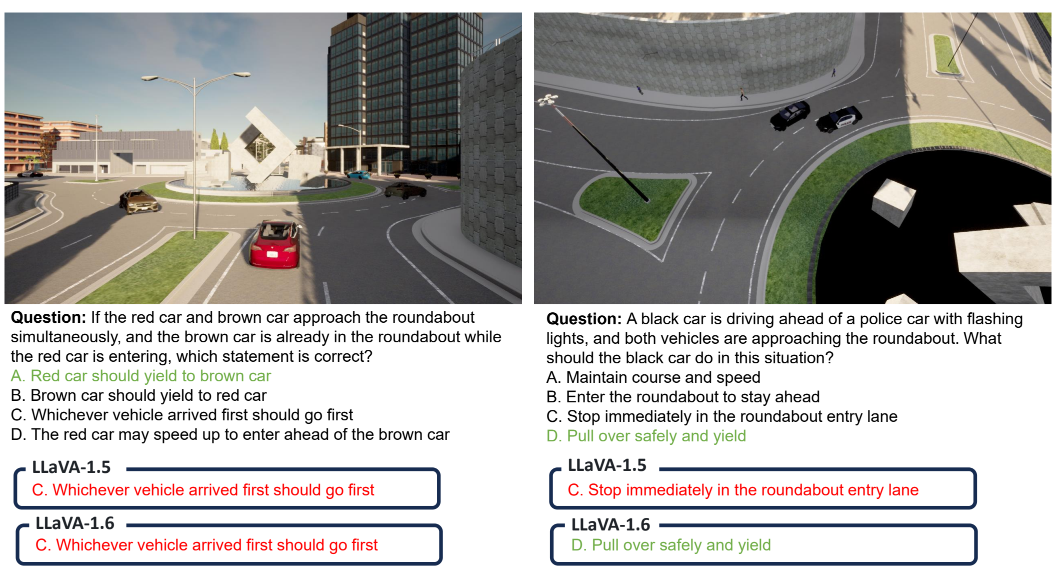


Figure S11. **Roundabout Q&A Examples and Model Predictions.** In green is the ground truth answers. In red is the error made by the model.

Question: From this driver's perspective, which traffic sign description is correct?

Horse

LLaVA

Trucks Out of Service. The image shows a sign that indicates the vehicle is a truck and is out of service.

Speed Limit 10 MPH

LLaVA

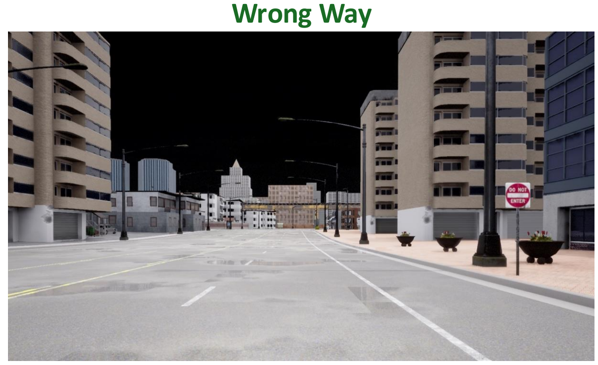
Speed Limit 15 MPH. The image shows a street sign with a red slash mark across a symbol, which indicates that the minimum speed limit is 15 MPH.

Slow

LLaVA

No Stopping. The image shows a city street at night with a no stopping sign.

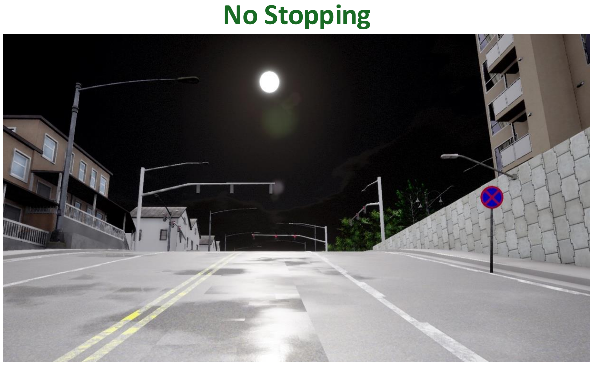
Wrong Way



LLaVA

No Stopping. The image shows a city street at night with a no stopping sign.

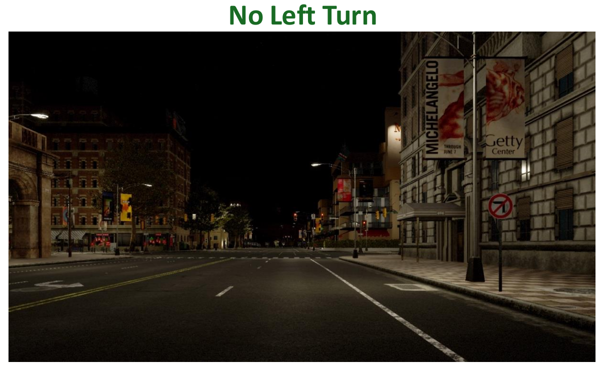
No Stopping



LLaVA

No Outlet Left. The STOP sign is handheld, which means that the person holding the sign must signal when it is safe to proceed.

No Left Turn



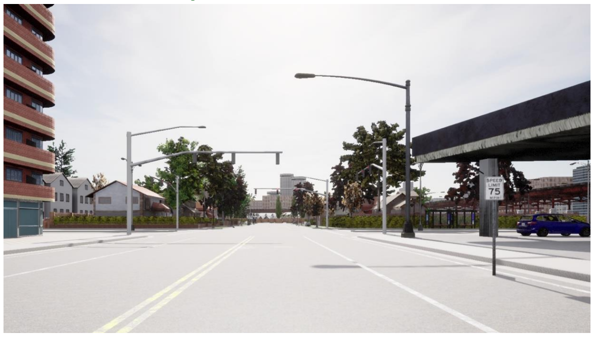
LLaVA

Chevron Left. The sign indicates that the driver must make a left turn.

Figure S12. **Traffic Sign Q&A Failure Cases (Part I).** Green captions are ground truth answers. In red is the error made by the model.

Question: From this driver's perspective, which traffic sign description is correct?

Speed Limit 75 MPH



VILA

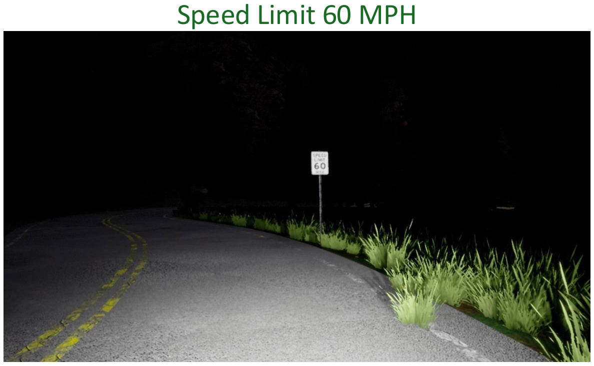
Road May Flood. The image shows a road that may flood.

No Bicycles

VILA

Stop Sign Ahead. The image shows a stop sign ahead, which is a common traffic sign used to indicate a complete stop before proceeding.

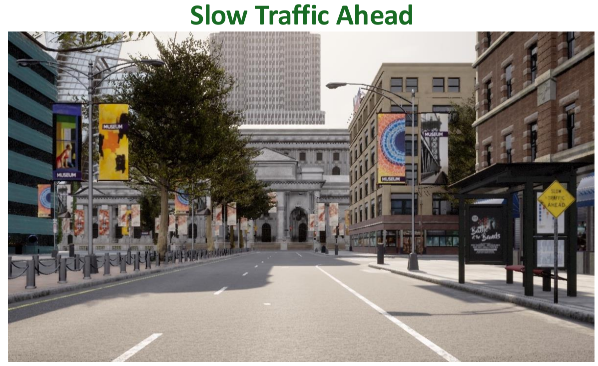
Speed Limit 60 MPH



VILA

One Lane Bridge. The sign indicates that there is a one lane bridge ahead.

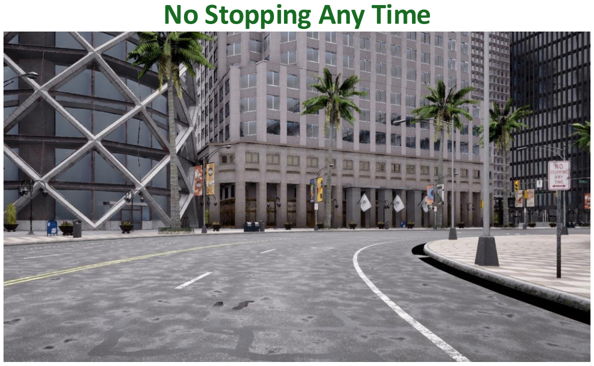
Slow Traffic Ahead



VILA

Yield. The image shows a city street with a yield sign. The sign is placed at an intersection, indicating that drivers must yield to other traffic before proceeding.

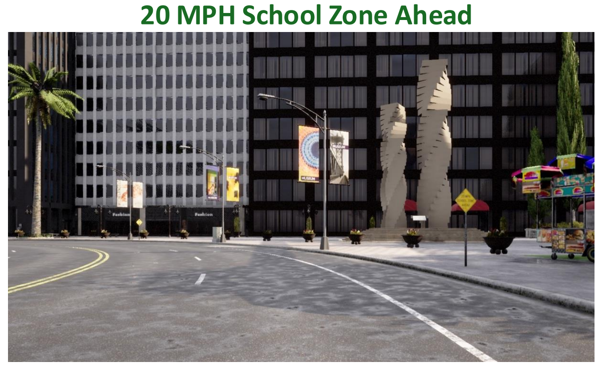
No Stopping Any Time



VILA

Yield. The image shows a city street with tall buildings and palm trees. The sign in the image is a yield sign, which means that drivers must come to a complete stop before proceeding.

20 MPH School Zone Ahead



VILA

Yield Sign. The sign is a white background with a red border and a red arrow pointing to the right.

Figure S13. **Traffic Sign Q&A Failure Cases (Part II).** Green captions are ground truth answers. In red is the error made by the model.

Question: From this driver's perspective, which traffic sign description is correct?

Truck Warning

GPT4o

Straight Ahead. The sign in the image is a yellow diamond with a black arrow pointing straight, which typically indicates a direction or path to follow.

Minimum Speed End 70 MPH

GPT4o

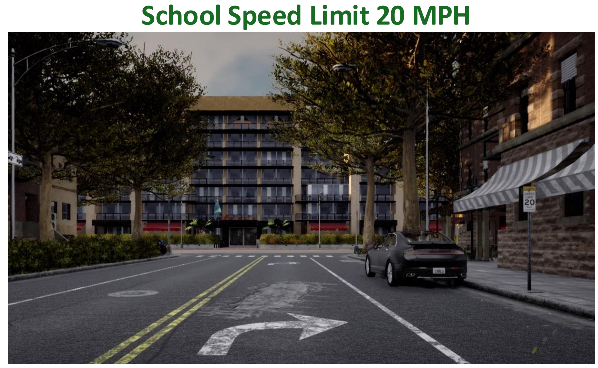
Minimum Speed 70 MPH. The sign indicates a minimum speed limit of 70 MPH. The blue circular sign typically indicates minimum speed in some regions.

Playground

GPT4o

A narrow bridge ahead. The arrows pointing inwards suggest that the road or bridge ahead is narrow.

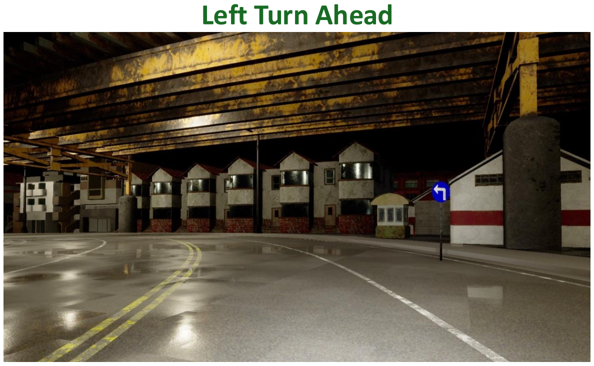
School Speed Limit 20 MPH



GPT4o

Left Arrow. The sign on the road indicates a left turn arrow.

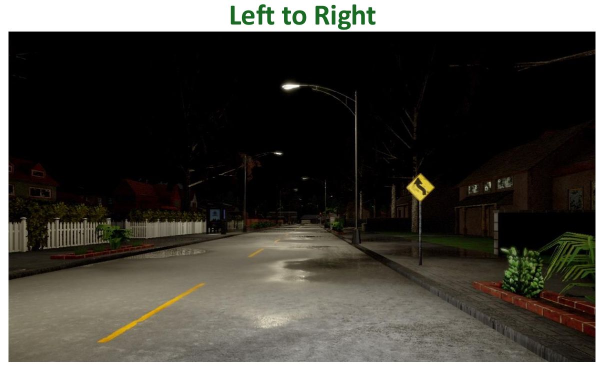
Left Turn Ahead



GPT4o

Right Arrow. The sign shows a right arrow, indicating a direction to the right.

Left to Right



GPT4o

Right Arrow. The sign in the image indicates a winding road ahead. It shows a curved arrow pointing to the right, which means the road curves in that direction.

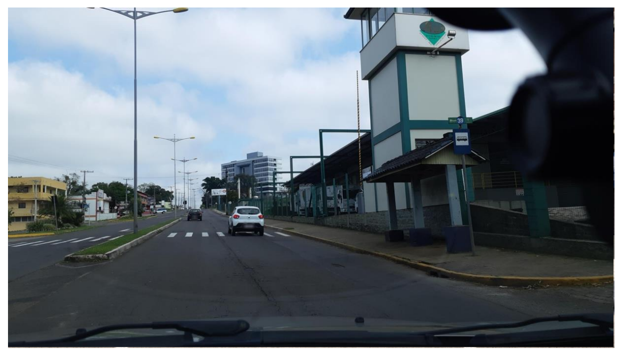
Figure S14. **Traffic Sign Q&A Failure Cases (Part III).** Green captions are ground truth answers. In red is the error made by the model.

Question: From this driver's perspective, which traffic sign description is correct?

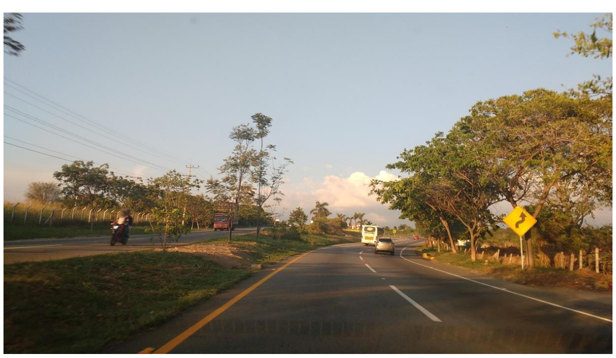
A. Maximum Speed Limit 70
B. Road Bump
C. Crossroads With Priority To The Right
D. End Of Speed Limit Zone

A. Bicycles Only
B. Parking
C. Pedestrians Crossing
D. Maximum Speed Limit 30

A. Pedestrians Crossing
B. Turn Left
C. Crossroads
D. No Motorcycles



A. Tram Bus Stop
B. Equestrians Crossing
C. T Roads
D. End Of Motorway



A. Double Curve First Right
B. Wild Animals
C. Stop Here On Red Or Flashing Light
D. Mopeds And Bicycles Only

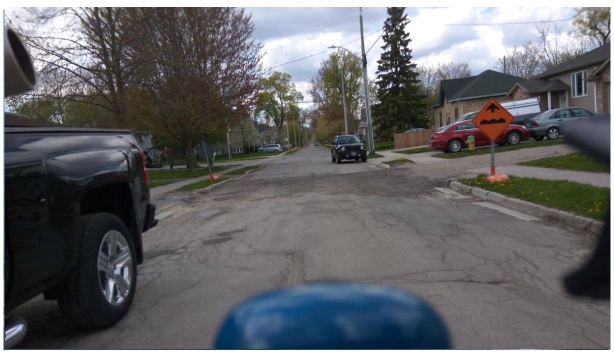
A. Stop
B. Tram Bus Stop
C. Uneven Roads Ahead
D. Maximum Speed Limit 60

Figure S15. **Traffic Sign Q&A Failure Cases on real world Mapillary dataset(Part I).** Green captions are ground truth answers, while in red is the error, made by the model LLaVA-1.5.

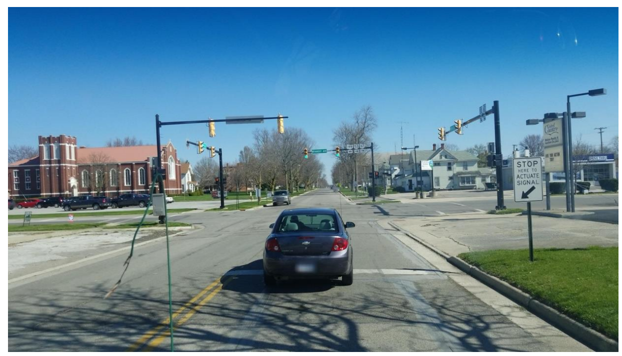
Question: From this driver's perspective, which traffic sign description is correct?

A. Maximum Speed Limit 65
B. Pass On Either Side
C. Passing Lane Ahead
D. End Of Speed Limit Zone

A. Railroad Crossing Without Barriers
B. Pedestrians Crossing
C. Horizontal Alignment Left
D. Domestic Animals



A. Curve Right
B. Uneven Roads Ahead
C. End Of Priority Road
D. Road Narrows Left



A. No Parking Or No Stopping
B. Stop Here On Red Or Flashing Light
C. Weight Limit
D. Stop Signals

A. Keep Right
B. Traffic Merges Left
C. Road Closed To Vehicles
D. End Of Pedestrians Only

A. Maximum Speed Limit 40
B. Hospital
C. End Of Priority Road
D. No U Turn

Figure S16. **Traffic Sign Q&A Failure Cases on real world Mapillary dataset(Part II).** Green captions are ground truth answers, while in red is the error, made by the model LLaVA-1.5.

Table S11. **Traffic Sign Recognition Accuracy for LLaVA-1.5.**

Sign Name	Accuracy (%)	Sign Name	Accuracy (%)	Sign Name	Accuracy (%)
End Detour	100.00	No Stopping Any Time Left	90.00	Straight Ahead	85.29
One Lane Bridge	100.00	No Standing Anytime	89.47	No Bicycles	85.19
End Road Work	100.00	Exit Open	89.47	No Parking	85.00
Narrow Bridge	100.00	Traffic Light Ahead	87.50	Yield Ahead	85.00
Road Work Next 5 Miles	100.00	Railroad	87.50	No Parking Any Time Left	84.85
Bump	100.00	Yield Ahead	87.50	Yield to Pedestrians	84.62
Detour	100.00	Priority Over Oncoming Traffic	87.50	Slippery Road	84.62
45 MPH Speed Zone Ahead	100.00	Handicap Bus Stop, No Standing	87.50	Double Arrow	84.62
Motorcycles	100.00	Speed Limit 60 MPH	87.50	No Parking	84.21
Exit Only	96.00	Left Arrow	86.96	Speed Limit 5 MPH	84.21
Exit Closed	95.00	Clearance 12'6"	86.67	Bicycles Keep Left, Pedestrians Keep Right	84.21
Left to Right	95.00	Traffic Light Ahead	86.67	No Train Horn	83.33
Pay Toll	94.12	Bike Lane	86.49	Deer	83.33
Yield	92.31	No Stopping Weekdays	86.36	Push Button	83.33
End	91.67	No Parking Any Time	86.36	Tow Away, No Stopping	83.33
Bend to Left	90.91	No Left Turn	86.36	Speed Limit 40 MPH	82.61
Straight Ahead	90.48	Right Turn	86.36	Downward Diagonal Left Arrow	82.61
Weight Limit	90.48	Stop	86.36	No Through Road	82.61
Minimum Speed 30 MPH	90.48	Signal Ahead	86.36	No Parking	82.61
Pedestrian	90.48	Left Turn Ahead	86.36	Speed Limit 50 MPH	82.35
Speed Limit 25 MPH	85.71	Speed Limit 20 MPH	82.35	Fog Area	82.14
No Standing Any Time	85.71	No Right Turn	82.35	Road Works	82.14
Trauma Center	85.71	No U-Turn Right	81.82	Curvy Road Right	81.82
Roundabout	85.71	Dead End Right	81.82	No Parking Except Sat, Sun, Holidays	81.48

Table S12. **Traffic Sign Recognition Accuracy for LLaVa-1.5 (Part II).**

Sign Name	Accuracy (%)	Sign Name	Accuracy (%)	Sign Name	Accuracy (%)
Right to Left	80.95	Minimum Speed End 40 MPH	80.00	Turn Right	78.57
Emergency Parking Only	80.95	Share the Road	80.00	White Circle, Blue Background, Red Strike	78.57
No Trucks	80.95	Slow	80.00	Speed Limit 55 MPH	78.26
Left Curve Ahead	80.95	Slow Traffic Ahead	78.26	Downward Diagonal Right Arrow	78.26
Uneven Road	80.65	Geese	77.78	Wheelchair	77.78
End School Speed Limit	80.00	Left Arrow	77.78	Danger	77.78
Left Turn	80.00	Right Turn Ahead	77.42	Right Arrow	77.27
Horse	77.27	No Standing, Cars Towed Away	77.27	Dead End Left	76.92
Supplemental Right Arrow	76.92	Chevron Left	76.47	Road Narrows	76.47
Curvy Road Left	76.47	Low Ground Clearance	76.47	Stop Here When Flashing	76.19
Height 14'4"	76.19	No Parking	76.19	No Parking Tow Zone	76.00
Minimum Speed End 50 MPH	75.00	End School Zone	75.00	Minimum Speed 70 MPH	75.00
No Train Horn	75.00	School Crossing	75.00	Work Zone	75.00
Speed Limit 30 MPH	75.00	Speed Limit 10 MPH	75.00	Monday - Friday	74.19
Right Arrow	74.19	Road Work	74.07	Emergency Stopping Only	74.07
Stop Here When Flashing	73.91	Minimum Speed 50 MPH	73.91	School Bus Stop Ahead	73.91
Pedestrian Crossing	73.68	Slippery Road	73.68	Chevron Right	73.68
No Outlet Right	73.33	Speed Limit 50 MPH	73.08	No Stopping Except on Shoulder	72.73
Wrong Way	72.73	Minimum Speed 40 MPH	72.41	Speed Limit 60 MPH	72.22
School Bus Turn Ahead	72.22	Speed Limit 15 MPH	72.00	Slow Traffic Ahead	72.00
Speed Limit 80 MPH	71.88	Stop Ahead	71.43	Crossroads with Priority	71.43
No Standing	71.43	Speed Limit 70 MPH	71.43	Clearance 3.8 m	70.97
Trucks Out of Service	70.59	Bicycles and Pedestrians	70.37	Wrong Way	70.37
Dead End	70.00	Bend to Right	69.70	Bus Stop	69.57
No Stopping Anytime	69.57	Plane 185	69.23	Slippery When Wet	68.97

Table S13. **Traffic Sign Recognition Accuracy for LLaVa-1.5 (Part III).**

Sign Name	**Accuracy (%)**	**Sign Name**	**Accuracy (%)**	**Sign Name**	**Accuracy (%)**
School Speed Limit 20 MPH When Flashing	68.75	Speed Hump	68.75	Playground	68.42
Speed Limit 90 MPH	68.42	Cross Only on Signal	68.42	No Parking Except Trucks	68.18
Minimum Speed End 30 MPH	68.18	No Standing Except Farmers Market	68.18	No U-Turns	68.18
No Stopping Any Time	68.00	Uneven Lanes	68.00	No Stops Tow Away Zone	68.00
Reverse Turn Left	67.86	Stop Ahead	67.86	No Stopping	66.67
Minimum Speed 60 MPH	66.67	No Parking Except Sat, Sun, Holidays	66.67	Road Narrows	66.67
No Stopping Except Trucks Loading and Unloading	66.67	Hospital	66.67	No Turns	66.67
Trail Crossing	66.67	Speed Limit 80 MPH	66.67	Double Arrow	66.67
Supplemental Left Arrow	66.67	20 MPH School Zone Ahead	66.67	Height 5 Feet	65.62
Speed Limit 30 MPH	65.38	Speed Limit 70 MPH	65.38	Road Hump	65.38
Minimum Speed End 70 MPH	65.22	Speed Limit 35 MPH	65.22	Traffic Circle	64.29
Speed Limit 75 MPH	64.29	Road May Flood	64.00	Speed Limit 45 MPH	64.00
No Standing Mon - Fri	64.00	Fire Truck	63.64	Truck Warning	63.64
Double Bend, First to Right	63.16	White Circle, Blue Background	63.16	Red Circle	62.96
Reverse Turn Right	61.90	Roundabout	61.90	Right Curve Ahead	61.90
Truck Warning	61.54	Minimum Speed End 60 MPH	61.54	Speed Limit 20 MPH	61.11
Golf Carts	60.87	Red Triangle	60.71	Road Works	60.00
No Stopping	60.00	No Parking on Pavement	60.00	Stop Sign Ahead	59.26
Speed Limit 65 MPH	59.09	No Outlet	58.82	Work Zone	58.33
Left Turn Ahead	57.89	No Passing Zone	57.89	No Pedestrians	57.14
Wrong Way	57.14	Clearance 12’6”	57.14	Weight Limit 2 Ton per Axle, 10 Tons Gross	57.14
No Traffic Signs	56.25	Double Bend, First to Left	55.56	No Outlet Left	50.00
Do Not Stop on Tracks	50.00	No Trucks Over 7000 LBS Empty WT	48.15	Tractor Crossing	45.45
School	41.67	Priority of Oncoming Traffic	40.47	7:30 AM - 8:30 AM, 2:30 PM - 3:30 PM	38.94

References

[1] 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. 4

[2] 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 *CVPR*, 2024. 7, 8

[3] Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv:2412.05271*, 2024. 5, 8

[4] Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. CARLA: An open urban driving simulator. In *CoRL*, 2017. 1

[5] 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. 4, 10

[6] Peng Gao, Jiaming Han, Renrui Zhang, Ziyi Lin, Shijie Geng, Aojun Zhou, Wei Zhang, Pan Lu, Conghui He, Xiangyu Yue, Hongsheng Li, and Yu Qiao. Llama-adapter v2: Parameter-efficient visual instruction model. *arXiv:2304.15010*, 2023. 8

[7] 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

[8] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*, 2019. 1

[9] Jinkyu Kim, Anna Rohrbach, Trevor Darrell, John Canny, and Zeynep Akata. Textual explanations for self-driving vehicles. In *ECCV*, 2018. 7, 8

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

[11] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *CVPR*, 2024. 5, 7, 8

[12] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, January 2024. 7

[13] Gerhard Neuhold, Tobias Ollmann, Samuel Rota Bulo, and Peter Kontschieder. The mapillary vistas dataset for semantic understanding of street scenes. In *ICCV*, 2017. 9

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

[15] 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. *arXiv:2312.14150*, 2023. 8

[16] 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. 4