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

Yue Li [1] Meng Tian [2] Zhenyu Lin [2] Jiangtong Zhu [2] Dechang Zhu [2]
Haiqiang Liu [2] Yueyi Zhang [1] Zhiwei Xiong [1,✉] Xinhai Zhao [2,✉]
[1] University of Science and Technology of China [2] Huawei Noah's Ark Lab
yueli65@mail.ustc.edu.cn zwxiong@ustc.edu.cn zhaoxinhai1@huawei.com

In the supplementary, we present more details about the benchmark, training setting, and results on each tasks.

Then we list the improvement rates of all tasks in *Traffic Graph Generation* aspect when the tips are added in instructions. Also, we compare the domain-specific model Dols [10] and the underlying foundational model Openflamingo [1] on static tasks.

Finally, we provide a detailed definition of each task in the proposed dataset **VLADBench**, illustrated with examples, including the images and corresponding questions. Note that the visual prompts and the choice lists are omitted.

1. Benchmark details

Following the selection principle, we first choose scenarios, annotate the visual elements with descriptions, design the questions, and then annotate the correct answers. 5 human annotators perform fine-grained annotations, with 2 researchers verifying the results. Each instance takes about 5 minutes to annotate.

The detailed numbers of tasks are listed as: Traffic Light (795 QAs), Pavement Marking (564 QAs), Traffic Sign (701 QAs), Right Of Way (309 QAs), VRU Recognition (424 QAs), Vehicle Recognition (223 QAs), Vehicle Status (257 QAs), Lane Recognition (780 QAs), Obstruction Recognition (680 QAs), Light (200 QAs), Weather (248 QAs), Sign-Sign Relation (192 QAs), Light-Lane Relation (702 QAs), Sign-Lane Relation (1072 QAs), Lane-Change Relation (784 QAs), Road-Speed Relation (340 QAs), VRU Cut-in (267 QAs), VRU Cross (276 QAs), Vehicle Cut-in (261 QAs), Long-Short Parking (320 QAs), VRU Bahavior (99 QAs), Vehicle Bahavior (80 QAs), Key Obstruction Detection (547 QAs), Risk Prediction (272 QAs), Drive Efficiency (303 QAs), Spatio-Temporal Reasoning (161 QAs), Lateral (235 QAs), Longitudinal (101 QAs), and Trajectory (799 QAs).

2. Training Setting

Training Data. The domain-specific dataset are collected from [3, 4, 7–17], including DriveLM-nuscenes [15] (377,956 QAs), LingoQA [13] (413,829 QAs), CODA-LM [4] (20,495 QAs), Dolphins [10] (102,025 QAs), IDKB [9] (188,486 QAs), MapLM [3] (143,252 QAs), DriveGPT4 [17] (26,751 QAs). Besides, we employ structured rules to generate 109,309 QAs, using the original annotations from [7, 8, 11, 14]. Then we use GPT-4o to increase the diversity of the 109,309 QAs.

For the trajectory training data, we selected 4,072 samples from nuScenes [2]. To generate analytical chain-of-thoughts (COT) for each scenario, we first collect questions related to each sample from [6, 15]. Then, we utilize GPT-4o to summarize these questions and transform them into declarative statements. Finally, by incorporating images, we use GPT-4o to refine the detailed scenario COT data.

Training Details. The training framework is inherited from IVL2 [5]. We finetune the pre-trained IVL2-4B with full parameters (including the vision encoder) for 2 epochs, with a batch size of 128 and a learning rate of 1e-5. The max token length is set to 4,096. All experiments are conducted on 16 nodes, each equipped with 8 V100 GPUs, with each task taking approximately 24 hours.

3. Bottlenecks in Traffic Graph Generation

As discussed, traffic graph generation is challenging, even with the provision of visual prompts. Besides, the descriptive guidance tips about traffic knowledge are incorporated into the instructions, including the meaning of signs, types of lights, and lane characteristics. To dig into underlying reasons, we calculate the accuracy improvement rate before and after adding these guidance tips. Tab. 1 and Tab. 2 list the accuracy improvement rate across the five tasks. Almost all models show improvement rate, suggesting that embedding traffic-related knowledge aids in graph construction.

For example, GPT-4o demonstrates a 42.54% improvement in sign-lane relation task and a 68.82% improvement in lane speed relation task. However, the final score for *signal element relation* remains far from 60, highlighting ongoing limitations in relational reasoning. The marginal improvement in the sign-sign relation task further underscores these constraints. Put things together, two points can be drawn: 1) Current VLMs still lack sufficient perception and understanding of traffic knowledge. 2) Even when provided with all the relevant knowledge, existing VLMs still exhibit weak reasoning capabilities in traffic graph tasks.

4. Dols versus Base Model

Dols [10] focuses on autonomous vehicle behavior understanding and achieves state-of-the-art performance in the vehicle cut-in intention judgment task. However, Dols [10] perform quite worse on most other tasks. One reason for this is the poor performance of the underlying foundation model (Openflamingo [1]), as demonstrated by the results on the static parts of our proposed dataset in Tab. 3. Additionally, we observe a notable performance drop in the *road traffic signals* aspect, with no improvements in recognition tasks or graph construction. These phenomenons highlight focusing on only one capability when adapting a foundational model to an autonomous driving domain is suboptimal. This approach prevents other relevant autonomous driving capabilities from improving and may even lead to significant performance degradation.

5. Detailed Results

The detailed scores of each tertiary tasks are listed in Tab. 5, and more motion planning results are presented in Tab. 4.

6. Examples of VLADBench

Figure 1. Examples of Pavement Marking.

Pavement Marking refers to markings painted or applied to roads, pavements, and other traffic areas to guide drivers, cyclists, and pedestrians. In this task, the questions include the type (a total of 20) and meaning of various pavement markings, including lane lines, arrows, symbols, text, etc., as illustrated by the samples in Fig. 1.

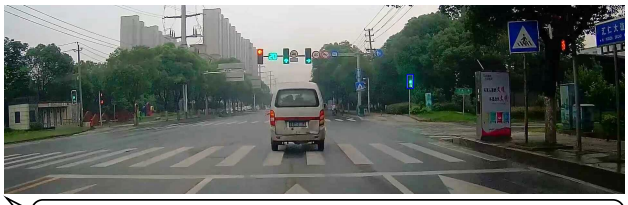
According to the directional sign within the red box in the image, where will the autonomous vehicle head if it turns left? [Choice List]

The lane sign within the red box in the image indicates a bus-only lane. Based on the image, could you please specify the designated operating hours for this lane? [Choice List]

Figure 2. Examples of Traffic Sign.

Traffic Sign refers to a visual display placed at the roadside or above a road to inform drivers, cyclists, and pedestrians about the road, its conditions, and traffic regulations. The category includes lane signs, directional signs, regulatory signs, prohibitory signs, warning signs and construction signs, encompassing a total of 168 types. Specifically, we present two important and intriguing questions in Fig. 2, which are designed to conduct an in-depth evaluation of VLM-based AD.

The state of the Motor Vehicle Light in the blue box is green. How many seconds are left?



The state of the Pedestrian Crossing Light in the blue box is green. How many seconds are left?

Figure 3. Examples of Traffic Light.

Traffic Light refers to a signaling device used to control the flow of traffic at intersections and other locations. The types

Table 1. The improvement rates of accuracy on traffic graph generation domain when the descriptive tips are added in the instructions. S.S. RL.: Sign-Sign Relation, S.L.RL.: Sign-Lane Relation, L.L.RL.: Light-Lane Relation, L.S. RL.: Lane Speed Relation, L.C. RL.: Lane Change Relation.

Task	VU 8B	IXC2.5 8B	VILA 8B	CV 8B	LoV 8B	QW 7B	MCV 8B	IVL2 8B	QW2 7B	OV 7B	QW2.5 7B	LV 7B	VILA 40B	OV 72B	LV 72B	IVL2 76B	QW2 72B	QW2.5 72B	GEM 1.5pro	GPT 4o
S.S.RL	2.08	1.04	3.13	0.00	1.04	1.04	14.58	3.13	0.00	0.00	1.04	1.04	3.13	4.17	4.17	0.00	0.00	3.13	12.5	20.83
S.L.RL	22.39	22.01	24.63	27.43	25.00	19.62	20.52	22.39	24.63	20.71	15.11	18.47	25.56	15.49	19.78	22.39	21.27	40.11	36.57	42.54
L.L.RL	25.93	19.94	11.68	44.44	21.08	26.50	25.64	30.20	21.94	21.08	23.65	11.97	15.67	50.43	52.42	32.19	35.61	34.19	43.02	32.76
MEAN	21.67	19.23	17.90	30.82	21.26	20.45	21.77	23.30	21.26	18.82	16.79	14.45	19.84	26.86	29.91	23.70	24.31	34.38	36.52	**36.93**
L.S.RL	18.24	57.06	19.41	24.71	8.24	1.18	35.88	36.47	50.59	32.94	57.06	35.88	31.76	48.24	48.24	63.53	71.18	68.24	51.76	68.82
L.C.RL	0.00	16.84	31.89	58.16	31.63	45.41	55.36	33.42	30.87	68.88	31.12	69.13	31.89	77.30	71.94	30.87	53.57	75.51	79.08	62.76
MEAN	5.52	29.00	28.11	48.04	24.56	32.03	49.47	34.34	36.83	58.01	38.97	59.07	31.85	68.51	64.77	40.75	58.90	**73.31**	70.82	64.59

Table 2. The improvement rate of accuracy on traffic graph generation domain. The results are from the domain-specific models.

Task	Dols 9B	Senna 7B	DriLM 4B	DriMM 7B	DriLM-B 4B	IVL2 4B	Ours 4B
S.S.RL.	3.13	0.00	3.13	7.29	0.00	0.00	2.08
S.L.RL.	16.79	1.68	5.60	27.43	16.98	11.38	27.99
L.L.RL.	17.38	0.28	37.32	18.80	29.34	15.95	34.76
MEAN	15.67	1.02	16.68	22.38	19.74	11.90	**27.87**
L.S.RL.	0.00	0.00	28.82	30.00	33.53	36.47	31.18
L.C.RL.	30.87	69.13	0.00	0.00	30.87	30.87	47.19
MEAN	21.53	48.22	8.72	9.07	31.67	32.56	42.35

include the motor vehicle light, non-motorized light, pedestrian crossing light, lane light, and arrow light. The traffic light statuses include red, yellow, green and malfunction. In addition to the type and status, we also incorporate questions regarding countdown timers, as illustrated in Fig. 3.

Figure 4. Examples of Right Of Way.

Right Of Way refers to the legal right of a person or vehicle to proceed before an ego vehicle at an intersection or other point on a road. Examples are illustrated in Fig. 4. Additionally, we provide descriptions of the actions to take when there are no visual prompts.

Figure 5. Examples of VRU Recognition.

VRU Recognition refers to the classification and detection for vulnerable road users (VRUs). There are 8 types of VRUs: moped, tricycle, cart, cyclist, bicycle, stroller, motorcycle, and wheelchair. Note that these VRUs are selected from corner cases. Examples are shown in Fig. 5.

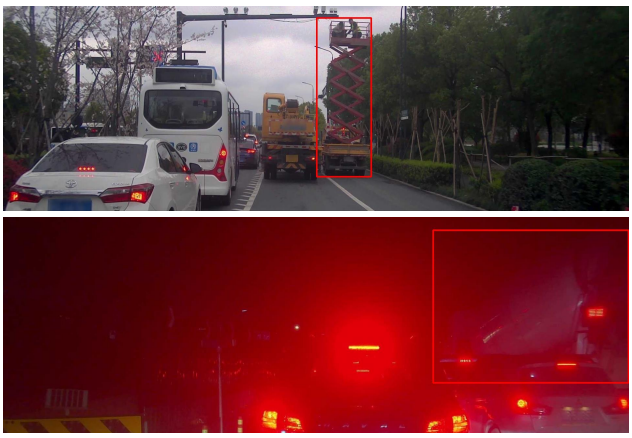


Figure 6. Examples of Vehicle Recognition.

Vehicle Recognition refers to the classification and detection for 5 types of corner-case level vehicles, including car, construction vehicle, truck, bus, and sanitation vehicle. Examples are shown in Fig. 6.

Table 3. Results on the static parts of our proposed dataset from the domain-specific model (Dols) and its foundation model (Openflamingo). P.M.: Pavement Marking, T.S.: Traffic Sign, T.L.: Traffic Light, LI.: Light, WE.: Weather, L.Rec.: Lane Recognition, V.S.: Vehicle Status, V.Rec.: Vehicle Recognition, VRU.Rec.: VRU Recognition, O.Rec.: Obstruction Recognition, S.S. RL.: Sign-Sign Relation, S.L.RL.: Sign-Lane Relation, L.L.RL.: Light-Lane Relation, L.S. RL.: Lane Speed Relation, L.C. RL.: Lane Change Relation, K.O.D: Key Object Detection.

Task	P.M.	T.S.	T.L.	LI.	WE.	L.Rec.	V.S.	V.Rec.	VRU.Rec	O.Rec.	S.S.RL.	S.L.RL.	L.L.RL.	L.S.RL	L.C.RL	K.O.D.
Openflamingo	30.92	31.61	58.04	45.2	54.84	46.21	27.32	0.00	0.38	0.00	1.04	2.10	5.84	1.18	7.46	0.00
Dolphins	12.09	25.02	42.92	56.40	60.65	30.97	3.97	0.00	0.00	0.00	0.78	4.20	4.34	0.00	7.72	74.11

Table 4. More motion planning results in DriveLM-nuScenes validation set from the State-of-the-art models and DS models.

Data	ST-P3								UniAD							
	L2(m)				Collision				L2(m)				Collision			
	1s	2s	3s	Avg	1s	2s	3s	Avg	1s	2s	3s	Avg	1s	2s	3s	Avg
ST-P3	1.28	2.03	2.81	2.04	0.14	0.72	1.28	0.71	-	-	-	-	-	-	-	-
Uniad	-	-	-	-	-	-	-	-	0.47	1.80	3.73	3.00	0.13	0.53	1.50	0.72
IVL-4B	5.93	7.39	8.91	7.41	6.51	8.49	9.73	8.25	6.92	9.58	12.66	9.72	7.66	10.44	14.00	10.70
DriveMM	11.46	15.05	18.74	15.08	1.92	6.46	12.93	7.12	10.24	12.17	14.06	12.06	1.06	2.94	5.69	3.23
GPT-4o	4.74	8.41	11.74	8.30	4.26	10.98	11.11	8.79	3.64	5.57	7.36	5.53	2.78	6.17	7.92	5.62
Gemini-1.5pro	3.70	6.67	9.44	6.61	5.11	8.68	7.54	7.11	2.78	4.35	5.84	4.33	3.32	5.62	6.45	5.13

Table 5. Detailed results evaluated on different VLMs. For the abbreviation, P.M.: Pavement Marking, T.S.: Traffic Sign, T.L.: Traffic Light, R.O.W: Right Of Way, LI.: Light, WE.: Weather, L.Rec.: Lane Recognition, V.S.: Vehicle Status, V.Rec.: Vehicle Recognition, VRU.Rec.: VRU Recognition, O.Rec.: Obstruction Recognition, S.S. RL.: Sign-Sign Relation, S.L.RL.: Sign-Lane Relation, L.L.RL.: Light-Lane Relation, L.S. RL.: Lane Speed Relation, L.C. RL.: Lane Change Relation, VRU.CI.: VRU Cut-in, V.CI: Vehicle Cut-in, VRU.C: VRU Cross, L.S.P.: Long-Short Parking, V.B.: Vehicle Behavior, VRU. B: VRU Behavior, K.O.D: Key Object Detection, ST.R.: Spatio-temporal Reasoning, R.P: Risk Prediction, D.E.: Drive Efficiency, LO: Longitudinal, LA: Lateral.

SubTask	VU 8B	IXC2.5 8B	VILA 8B	CV 8B	LoV 8B	QW 7B	IVL2 4B	MCV 8B	IVL2 8B	QW2 7B	OV 7B	QW2.5 7B	LV 7B	VILA 40B	OV 72B	LV 72B	IVL2 76B	QW2 72B	GEM. 1.5pro	GPT 4o	QW2.5 72B	Dols 9B	DriLM 4B	DriMM 7B	DriLM-B 4B	Ours 4B
T.L.	15.17	30.24	44.53	62.82	44.88	67.37	69.51	60.86	67.70	65.11	62.57	62.77	60.05	43.85	73.11	70.92	71.32	66.79	66.79	67.72	74.54	42.92	75.14	68.20	71.77	74.94
P.M.	34.79	22.59	31.21	38.51	50.60	36.42	38.62	34.72	46.38	57.38	58.37	62.91	63.51	31.03	53.72	65.11	51.38	66.88	73.01	71.17	72.45	12.09	39.50	54.11	40.99	58.33
T.S.	26.05	20.68	13.50	35.89	18.46	33.58	34.01	32.10	47.45	40.94	49.27	61.71	49.73	20.29	32.98	40.83	53.92	71.18	64.05	68.96	65.11	25.02	44.74	47.05	40.09	61.00
MEAN	24.24	24.89	30.32	47.00	37.46	47.40	48.97	43.91	54.97	54.77	56.89	62.45	57.49	32.32	54.15	59.09	59.94	68.31	67.56	69.09	70.76	28.39	55.04	57.15	52.56	65.65
R.O.W.	33.92	32.69	46.34	59.35	49.45	79.22	80.58	22.72	69.45	71.52	70.81	80.32	74.11	46.34	76.96	80.06	71.07	81.36	42.98	78.96	80.58	21.10	81.36	42.33	73.85	80.58
MEAN	33.92	32.69	46.34	59.35	49.45	79.22	80.58	22.72	69.45	71.52	70.81	80.32	74.11	46.34	76.96	80.06	71.07	81.36	42.98	78.96	80.58	21.10	81.36	42.33	73.85	80.58
VRU.RG	9.32	20.78	9.31	20.83	21.21	18.76	39.10	34.70	38.34	45.24	38.58	39.79	35.96	25.47	33.42	32.61	44.94	53.12	43.73	40.58	38.85	21.20	30.47	36.30	34.56	46.05
V.RG	36.16	42.19	13.27	27.33	36.39	50.51	61.33	52.51	61.92	66.07	63.39	52.54	58.27	41.99	57.75	58.20	64.33	64.70	67.61	58.75	58.78	46.41	47.71	58.74	56.76	74.14
V.S.	23.50	23.66	26.77	24.67	41.95	24.36	51.05	54.09	49.26	47.16	54.09	52.61	52.53	48.40	53.62	54.94	59.69	54.16	55.41	54.86	55.95	3.97	51.36	55.49	48.79	56.96
L.RG	22.36	32.92	45.97	54.36	49.26	54.23	58.26	47.90	48.62	52.51	67.69	63.49	66.15	49.85	67.54	64.05	64.44	62.87	51.26	63.49	70.87	30.97	73.44	66.49	66.62	74.26
O.RG	12.80	17.58	26.77	7.59	35.15	32.87	49.83	39.81	52.94	48.95	44.77	51.74	47.11	31.50	39.88	44.28	46.96	63.62	50.24	48.97	50.51	36.18	45.25	46.07	44.65	50.47
MEAN	18.70	26.20	28.70	29.11	38.16	38.13	51.90	44.31	49.34	50.88	53.99	53.64	53.04	39.30	51.03	51.18	55.39	60.56	51.61	53.82	56.51	29.24	52.80	53.27	51.68	60.47
LI.	45.80	16.40	63.80	58.60	62.50	58.30	63.30	70.40	66.30	70.40	69.60	68.40	70.00	62.20	70.00	70.40	70.40	67.20	56.40	67.20	73.20	56.40	63.60	68.70	66.40	68.40
WE.	54.11	27.50	65.81	58.95	65.08	68.55	69.68	71.05	73.87	71.61	68.71	70.00	68.39	67.50	71.21	72.26	74.52	72.26	73.23	69.27	70.97	60.65	65.16	71.61	69.68	74.19
MEAN	50.40	22.54	64.91	58.79	63.93	63.97	66.83	70.76	70.49	71.07	69.11	69.29	69.11	65.13	70.67	71.43	72.68	70.00	65.71	68.35	71.96	58.75	64.46	70.31	68.21	71.61
S.S.RL	2.08	0.49	21.80	10.00	21.02	17.16	19.90	29.69	21.72	20.00	22.00	20.96	23.22	21.28	22.20	23.90	20.00	20.13	28.18	36.56	22.15	18.28	9.71	24.84	20.00	19.71
L.L.RL	22.39	15.33	34.70	32.17	37.76	25.56	32.73	37.35	41.56	38.38	37.35	37.17	34.83	35.39	50.79	50.54	47.05	41.84	50.80	50.27	51.34	26.45	36.49	32.73	41.16	57.01
S.L.RL	25.93	16.77	39.03	24.81	39.44	23.54	28.35	34.29	36.87	39.02	35.23	31.61	35.08	40.00	30.02	33.14	37.21	33.88	43.44	51.41	50.21	27.64	22.40	40.16	33.66	46.50
MEAN	21.67	14.67	35.80	25.99	37.04	23.64	29.09	34.93	37.07	36.93	34.70	32.56	33.83	36.52	36.68	38.45	39.04	35.38	44.58	49.56	47.88	26.08	26.19	36.01	35.00	47.64
L.C.RL	18.24	26.43	48.57	69.27	47.47	52.69	44.76	63.33	46.48	44.69	72.49	44.77	75.31	47.42	72.65	77.23	44.69	59.86	73.95	63.70	71.61	0.00	13.88	16.89	53.69	58.33
L.S.RL	0.00	59.21	12.96	19.04	34.82	0.93	47.93	44.82	48.15	55.84	46.40	58.44	47.53	44.60	53.96	54.18	60.53	70.91	62.29	70.35	70.54	44.63	38.04	39.18	45.43	44.06
MEAN	5.52	36.34	37.80	54.08	43.65	37.03	45.72	57.73	46.98	48.06	64.60	48.91	66.90	46.57	67.00	70.26	49.48	63.20	70.43	65.71	71.29	31.13	21.19	23.63	51.19	54.01
VRU.CI	12.00	61.39	44.64	57.83	46.37	62.40	55.51	57.36	62.53	61.07	59.96	53.76	59.66	45.84	62.38	62.38	63.31	68.54	64.72	54.96	59.38	51.25	58.65	58.41	63.30	63.43
VRU.C	5.71	88.15	35.94	31.16	37.55	80.92	37.23	78.86	45.49	82.52	74.96	70.98	77.77	37.52	66.20	55.49	65.05	75.38	72.39	51.07	54.46	66.05	44.29	47.45	83.97	59.40
V.C	12.19	62.59	36.25	20.80	33.05	83.10	21.84	52.24	24.14	39.31	36.17	25.52	42.53	36.51	27.64	28.47	26.93	34.48	26.17	23.54	28.28	86.70	29.31	18.72	44.60	26.09
L.S.P	19.19	58.69	24.44	44.81	24.06	50.00	49.75	52.56	53.25	58.75	58.13	60.75	57.50	24.56	45.00	54.00	57.25	62.75	58.75	58.75	71.00	32.00	54.50	47.06	52.25	60.63
MEAN	12.63	67.47	34.80	38.98	34.76	68.23	41.56	60.08	46.79	60.62	57.60	53.42	59.52	35.57	50.30	50.43	53.57	60.66	55.95	47.79	54.26	57.64	47.13	43.27	60.89	52.97
VRU.B.	16.25	27.88	10.50	43.64	23.84	29.90	43.43	41.21	45.86	44.04	39.80	44.04	42.42	11.50	45.25	46.87	47.88	44.24	53.94	58.79	52.93	0.25	36.77	40.20	42.02	41.21
V.B.	18.99	34.00	10.50	21.50	10.75	27.00	45.00	33.00	48.00	39.00	44.00	40.00	44.00	11.50	37.00	41.25	47.00	47.00	46.50	45.00	47.00	0.00	45.75	40.00	44.00	45.00
MEAN	17.77	30.61	10.50	33.74	17.99	28.60	44.13	37.54	46.82	41.79	41.68	42.23	43.13	11.50	41.56	44.36	47.49	45.47	50.61	52.63	50.28	0.11	40.78	40.11	42.91	42.91
K.O.D	0.00	26.62	54.52	44.42	53.93	76.01	54.66	59.49	69.73	62.41	59.49	58.17	69.73	55.50	71.33	69.87	73.53	72.94	63.44	70.60	75.72	74.11	66.95	64.75	75.72	74.55
R.P.	13.36	73.09	75.06	11.94	69.47	62.16	31.18	72.61	59.29	41.91	30.88	55.40	33.04	71.75	49.26	42.32	73.67	44.92	57.10	66.64	55.18	47.68	40.88	24.78	33.60	79.15
D.E.	50.88	54.92	38.88	36.44	38.94	44.95	46.47	48.78	54.85	51.68	32.81	61.19	53.99	40.86	48.38	50.83	56.30	63.04	60.33	62.71	63.70	68.01	53.47	59.27	46.60	58.15
ST.R	29.24	48.58	33.98	54.30	40.75	64.08	44.49	52.36	53.06	57.59	61.29	62.72	55.92	34.42	32.27	50.95	46.36	71.56	62.18	53.61	65.71	28.45	48.73	48.82	49.26	59.20
MEAN	19.37	45.91	52.60	36.89	52.03	64.24	46.47	58.85	61.91	54.93	47.35	58.87	56.50	52.84	56.33	57.16	66.08	64.49	61.20	65.77	67.27	58.72	55.95	52.99	56.60	69.73
LA.	6.73	63.91	22.55	24.17	23.49	26.81	56.77	40.26	41.45	30.21	32.43	33.62	41.11	24.43	46.81	35.15	62.55	55.40	56.09	43.83	44.85	18.22	39.74	50.98	53.36	59.83
LO.	27.74	36.24	9.31	20.20	6.73	17.03	26.34	24.95	39.41	46.93	46.34	39.80	41.39	7.33	68.91	46.14	48.51	50.10	57.23	57.82	54.85	11.74	31.49	47.72	29.90	50.89
MEAN	21.43	55.60	18.57	22.98	18.45	23.87	47.62	35.65	40.83	35.24	36.61	35.48	41.19	19.29	53.45	38.45	58.33	53.81	56.43	48.04	47.86	13.69	37.26	50.00	46.31	57.14
Overall	20.14	32.34	35.85	38.16	40.06	44.78	46.21	47.28	50.27	51.07	52.15	52.30	53.63	39.62	52.64	54.04	54.89	58.80	57.19	58.92	61.03	34.97	45.75	47.01	51.32	59.45

Figure 7. Examples of Obstruction Recognition.

Obstruction Recognition refers to the classification and detection for the 19 types of corner-case level obstacles, including debris, suitcase, concrete block, plastic bag, chair, machinery, phone booth, dustbin, basket, stone, garbage, tire, carton, cardboard, garbage bag, traffic cone, traffic box, traffic island, and dog. These obstacles, often found unexpectedly on the road or in parking areas, pose significant risks to the safety and efficiency of autonomous vehicles. The ability to accurately detect and classify such obstructions is critical for ensuring smooth navigation and avoiding accidents in dynamic environments. Examples are shown in Fig. 7.

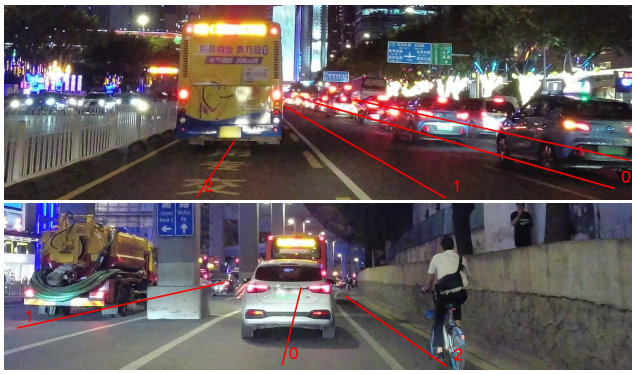


Figure 8. Examples of Lane Recognition.

Lane Recognition involves identifying various types of lanes, such as motorized vehicle lanes, non-motorized vehicle lanes, emergency lanes, dedicated bus lanes, ETC-exclusive lanes, and sidewalks. Additionally, we increase the complexity of the lane recognition task by distinguishing between opposing lanes and classifying motor vehicle lanes as fast, slow, or regular. The added complexity helps validate the ability to navigate complex roadways, enhancing its decision-making process in various traffic conditions. Accurate lane recognition is crucial for safe lane changes, merging, and ensuring the vehicle stays within the correct lane. Examples are shown in Fig. 8.

Vehicle Status focuses on the light state and operational status, referring to observable exterior conditions of a target vehicle. There are 17 possible statuses, including right turn light on, left turn light on, brake light on, hazard lights on, empty car light on, passenger light on, right door open, left door open, trunk open, trunk open for loading, right door open for boarding, left door open for boarding, rear compartment door open for loading, all compartment doors open, construction work, accident scene, and cargo hanging from trunk. The exterior status provides critical information about the current state of the target vehicle, which is essential for autonomous systems to make accurate decisions in real-time traffic scenarios. For instance, recognizing whether doors are open or if hazard lights are on helps the system assess the vehicle's intent or possible hazards. Examples are shown in Fig. 9.

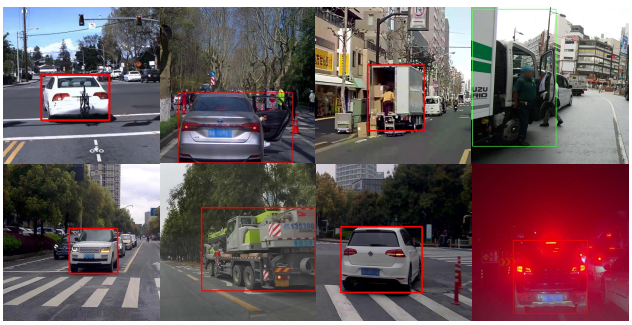


Figure 9. Examples of Vehicle Status.

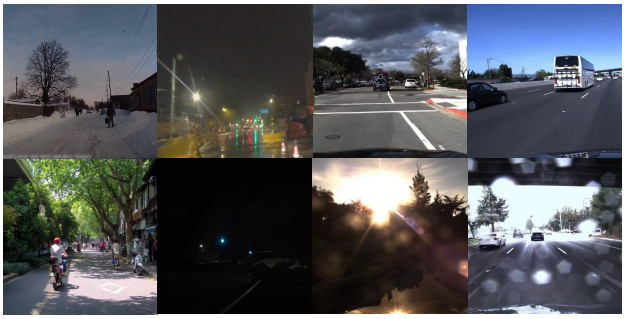
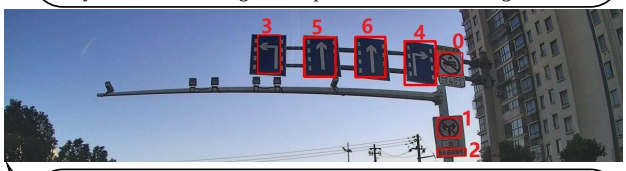


Figure 10. Examples of Weather & Light .

Weather & Light refer to the driving conditions in the scenarios. The questions are first addressed in terms of daytime, nighttime, and dawn&dusk conditions, followed by considerations of light and weather. The light conditions comprise diffuse, backlit, snowy, shadowed light, bright, low, very low, and dark. The weather conditions consist of overcast, clear, rainy, cloudy, and snowy. These environmental factors significantly impact the perception and decision-making processes of AD systems, as they affect visibility, road conditions, and overall driving safety. For instance, driving in low-light or snowy conditions requires the vehicle to adjust its speed and navigation strategy. Examples are shown in Fig. 10.

Sign-Sign Relation refers to the traffic graph connections between different signs. In the image, traffic signs are highlighted with red boxes, and the closest red numbers indicate

1. Analyze which traffic sign corresponds to No.5 traffic sign.
2. No.0 traffic sign is No entry for trucks sign, No.1 traffic sign is Straight ahead on second lane sign, No.2 traffic sign is Includes special operation vehicle sign, No.3 traffic sign is Straight ahead on first lane sign, No.4 traffic sign is Straight ahead on third lane sign, No.5 traffic sign is Construction vehicles use second lane sign. Analyze which traffic sign corresponds to No.5 traffic sign.



1. Analyze which traffic sign corresponds to No.0 traffic sign.
2. No.0 traffic sign is 00:00-24:00, no right turns for trucks over X tons sign, No.1 traffic sign is No left or right turn sign, No.2 traffic sign is Hazardous materials transport vehicle sign, No.3 traffic sign is Left turn sign, No.4 traffic sign is Right turn sign, No.5 traffic sign is Go straight sign, No.6 traffic sign is Go straight sign. Analyze which traffic sign corresponds to No.0 traffic sign.

Figure 11. Examples of Sign-Sign Relation.

the corresponding traffic sign identifiers. The Sign-Sign Relation task analyzes which sign corresponds to a specified sub-sign, capturing the hierarchical or contextual relationships between signs. Examples are shown in Fig. 11.

Figure 12. Examples of Lane-Sign Relation.

Lane-Sign Relation refers to the traffic graph about the lanes and traffic signs. In the image, the lanes are represented by red lines, with the adjacent red numbers indicating the lane identification. Traffic signs are depicted within blue boxes, with the nearest blue numbers indicating traffic sign numbers. The Lane-Sign Relation task focuses on determining which lane corresponds to a specified traffic sign, enabling the system to interpret lane-specific rules or guidance effectively. Examples are shown in Fig. 12.

Light-Lane Relation refers to the traffic graph about the lanes and lights. In the image, the lanes are represented by red lines, with the adjacent red numbers indicating the lane identification. Traffic lights are represented by blue boxes, with the adjacent blue numbers indicating the traffic light

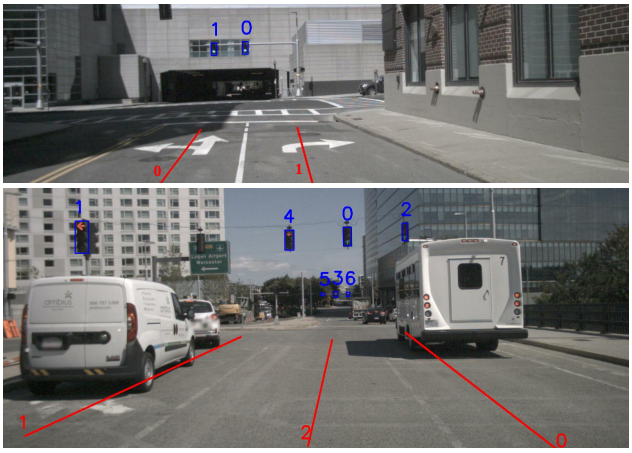


Figure 13. Examples of Light-Lane Relation.

identification. The Light-Lane Relation task analyzes which lane corresponds to a specified traffic light, facilitating an understanding of how traffic signals regulate specific lanes. Examples are shown in Fig. 13.

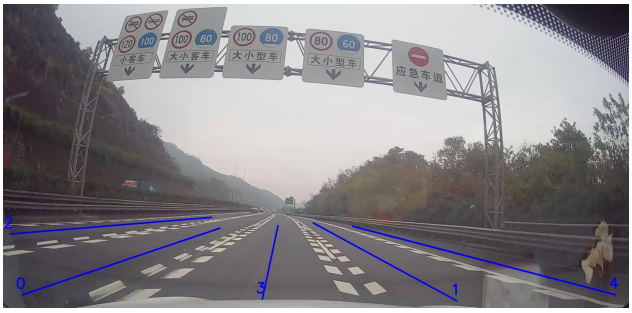


Figure 14. Examples of Lane Speed Relation.

Lane Speed Relation involves analyzing the low-speed and high-speed limits for a specified lane, based on the traffic signs present in the image. This task requires identifying the speed-related traffic signs and associating them with the relevant lanes to determine the permissible speed range. Examples are shown in Fig. 14.

Lane Change Relation analyzes the permissibility and rules governing lane changes. This task is based on road markings, represented here by a blue box, and involves determining whether a lane change from one lane to another is allowed. Understanding lane-change relationships is critical for autonomous systems to safely navigate dynamic traffic environments, such as highways or multi-lane roads, where precise adherence to road markings is necessary to avoid collisions and ensure smooth traffic flow. Examples are shown in Fig. 15.

Vehicle Cut-in refers to the task of judging whether a target vehicle intends to merge from an adjacent lane or other areas into the lane of the ego vehicle, and analyzing the motivation behind the behavior. Examples are shown in Fig. 16.

VRU Cut-in refers to the task of judging whether a target

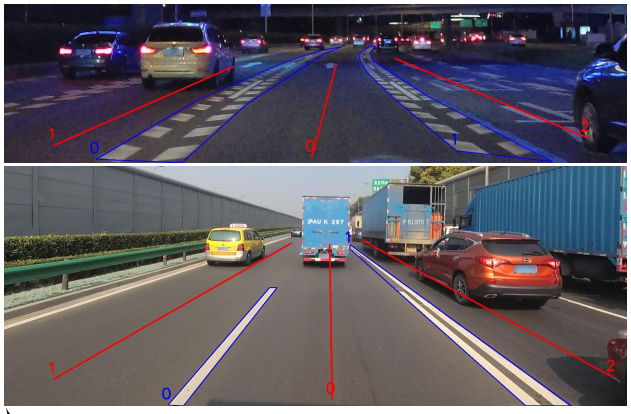


Figure 15. Examples of Lane Change Relation.

Figure 16. Examples of Vehicle Cut-in.

Figure 17. Examples of VRU Cut-in.

VRU intends to merge from a different lane into the lane of the ego vehicle, and analyzing the motivation behind the behavior. Examples are shown in Fig. 17.

Figure 18. Examples of VRU Cross.

VRU Cross refers to determining whether a VRU intends to cross laterally from one side to the other across the ego vehicle's path of travel, and analyzing the motivation behind the behavior. Examples are shown in Fig. 18.

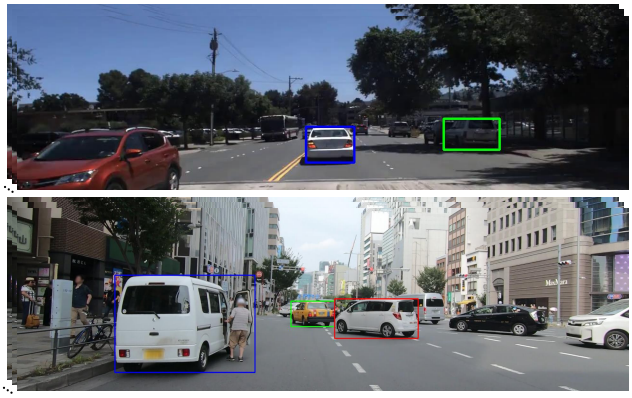


Figure 19. Examples of Long-Short Parking.

Long-Short Parking focuses on analyzing the parking time of the target vehicle. Whether the target vehicle is considered to be long-term or short-term parking (e.g., waiting for the traffic light, yielding, ever-changing passengers) is determined by whether the ego vehicle needs to perform a detour or execute an escape maneuver. Examples are shown in Fig. 19.

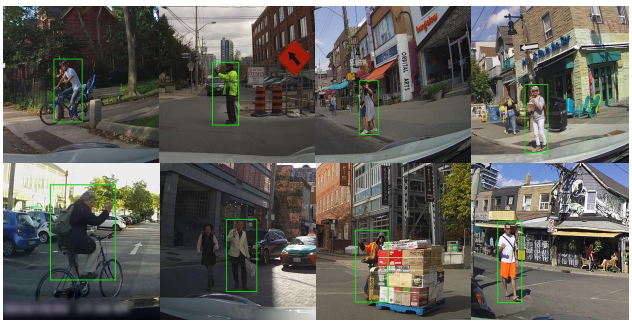


Figure 20. Examples of the pedestrian gesture of VRU Behavior.

Vehicle & VRU Behavior focuses on describing events that

have occurred, with an emphasis on understanding and interpreting behaviors in the traffic environment. Vehicle behavior is characterized by longitudinal and lateral movements, capturing actions such as acceleration, braking, and lane changes. VRU behavior encompasses critical maneuvers, including cut-in and crossing actions, which are essential for predicting potential conflicts. Additionally, pedestrian gesture analysis is incorporated to assess claims of right of way, providing a deeper evaluation of interactions between pedestrians and vehicles. Examples are shown in Fig. 20.

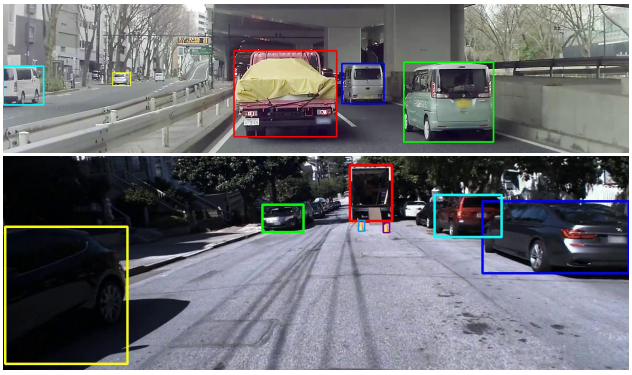


Figure 21. Examples of Key Object Detection.

Key Object Detection refers to the identification of objects that play a critical role in determining the vehicle's ability to maintain its current trajectory or safely execute lane changes to the left or right. These key objects may include vehicles, obstacles, or environmental elements that directly or indirectly influence driving decisions and safety. Examples are shown in Fig. 21.

Drive Efficiency aims to evaluate the operational effectiveness of ego vehicles in relation to traffic congestion levels. The evaluation framework considers three aspects: the current driving efficiency under prevailing congestion conditions, projected efficiency changes as congestion evolves, and the factors influencing these changes. By emphasizing congestion as a key metric, this task provides insights into optimizing driving strategies in dense traffic environments. Examples are shown in Fig. 22.

Risk Prediction evaluates the presence of significant potential risks in the environment as the vehicle proceeds straight or attempts to change lanes to the left or right. This task is divided into two steps: first, determining whether a risk exists; and second, given the source of the risk, analyzing the underlying cause of the risk. Examples are shown in Fig. 23.

Spatio-Temporal Relation leverages information from preceding frames to infer the current driving conditions, such as unseen or occluded traffic lights and signs, or the attributes of lanes. The questions in this task are designed to require associative reasoning or recollection of previously observed information, instead of being solvable directly through simple visual cues. Examples are shown in Fig. 24.

1. In the given autonomous driving image sequence, at the moment of the final image, how about the driving efficiency of the ego vehicle? [Choice List].
2. In the given autonomous driving image sequence, the ego vehicle is moving straight, at the moment of the last image, how the driving efficiency of the ego vehicle will be? [Choice List]
3. In the given autonomous driving image sequence, the ego vehicle is moving straight, at the moment of the last image, the future driving efficiency of the ego vehicle will increase. Select the most appropriate reason from the following options: [Choice List]

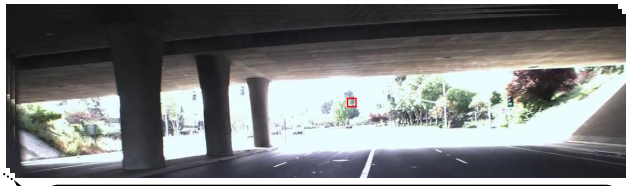
Figure 22. Examples of Drive Efficiency.

1. Based on the given sequence of images, assess whether there are any significant potential risks present in the environment if the vehicle proceeds straight or changes lanes to the left or right.
2. In the given image sequence, the pedestrian poses significant risks to the vehicle's straight driving and lane changes to the left or right. Please choose the most appropriate description to describe this risk:[Choice List]
3. In the given image sequence, the cyclist poses significant risks to the vehicle's straight driving and lane changes to the left or right. Please choose the appropriate description for the risk:[Choice List]

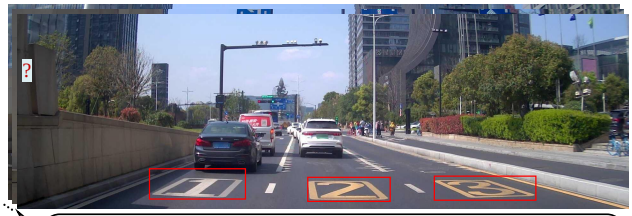
Figure 23. Examples of Risk Prediction.

Longitudinal refers to the management of a vehicle's speed and acceleration/deceleration along its direction of travel. The longitudinal operation includes maintain speed, accelerate, stop, decelerate, and decelerate to stop. Examples are shown in Fig. 25.

Lateral refers to the management of a vehicle's position

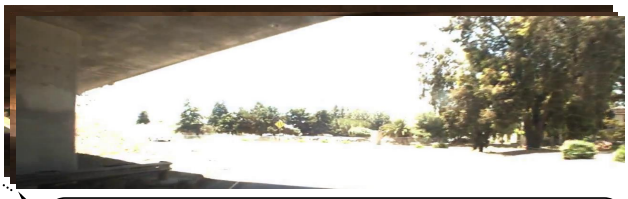


1. In the given autonomous driving image sequence, at the moment of the final image, which type is the ego lane?
2. In the given autonomous driving image sequence, at the moment of the final image, what is the status of the traffic light at this time?

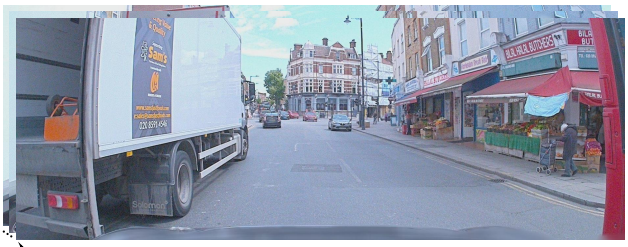


1. In the given autonomous driving image sequence, at the moment of the final image, what is the type of the lane ego vehicle in?
2. In the given autonomous driving image sequence, at the moment of the final image, is the right lane available for going straight?

Figure 24. Examples of Spatio-Temporal Relation.



Based on the given sequence of autonomous driving vehicle, the current environment is exiting a tunnel ahead, backlight causes inability to observe ahead, ego vehicle is going straight. Based on the assessment of driving risk and traffic efficiency, what longitudinal decision should the vehicle make at this moment?

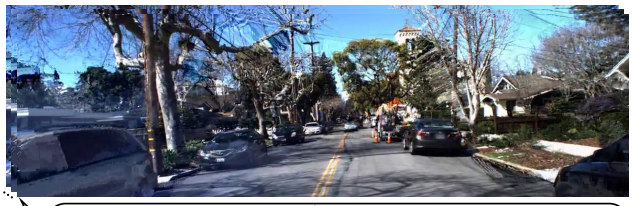


Based on the given sequence of autonomous driving vehicle, the current environment is the road ahead is narrow, an oncoming vehicle is a bus, and about to meet, ego vehicle is going straight. Based on the assessment of driving risk and traffic efficiency, what longitudinal decision should the vehicle make at this moment?

Figure 25. Examples of Longitudinal.

and direction within its lane or on the road. The lateral operation includes in-lane left avoidance, in-lane right avoidance, maintain straight or change lane to the left, maintain straight or change lane to the right, borrow lane for right avoidance, borrow lane for left avoidance, change lane to the right, change lane to the left, change lane to the left or right, and maintain straight. Examples are shown in Fig. 26.

Trajectory prediction is formulated as a vision-language task, incorporating critical perception and prediction results



Based on the given sequence of autonomous driving vehicle, the current environment is narrow road, ahead right stopped construction vehicle, ego vehicle proceeds straight. Based on the assessment of driving risk and traffic efficiency, what lateral decision should the vehicle make at this moment?

1. Ego vehicle proceeds straight through. Based on the assessment of driving risk and traffic efficiency, what lateral decision should the vehicle make at this moment?
2. Ego vehicle turns left. Based on the assessment of driving risk and traffic efficiency, what lateral decision should the vehicle make at this moment?
3. Ego vehicle turns right. Based on the assessment of driving risk and traffic efficiency, what lateral decision should the vehicle make at this moment?

Figure 26. Examples of Lateral.

along with high-level decisions. Besides, the ego status and the historical waypoints (last 2 seconds, given by four points) are included in the instruction. The VLMs then generate a feasible 3-second driving trajectory, consisting of 6 waypoints. An example is shown in Fig. 27.

References

[1] Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Shiori Sagawa, et al. Openflamingo: An open-source framework for training large autoregressive vision-language models. *arXiv preprint arXiv:2308.01390*, 2023. 1, 2

[2] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *Conference on Computer Vision and Pattern Recognition*, 2020. 1

[3] Xu Cao, Tong Zhou, Yunsheng Ma, Wenqian Ye, Can Cui, Kun Tang, Zhipeng Cao, Kaizhao Liang, Ziran Wang, James M Rehg, et al. Maplm: A real-world large-scale vision-language benchmark for map and traffic scene understanding. In *Conference on Computer Vision and Pattern Recognition*, pages 21819–21830, 2024. 1

[4] Kai Chen, Yanze Li, Wenhua Zhang, Yanxin Liu, Pengxiang Li, Ruiyuan Gao, Lanqing Hong, Meng Tian, Xinhai Zhao, Zhenguo Li, et al. Automated evaluation of large vision-language models on self-driving corner cases. *arXiv preprint arXiv:2404.10595*, 2024. 1

[5] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu,

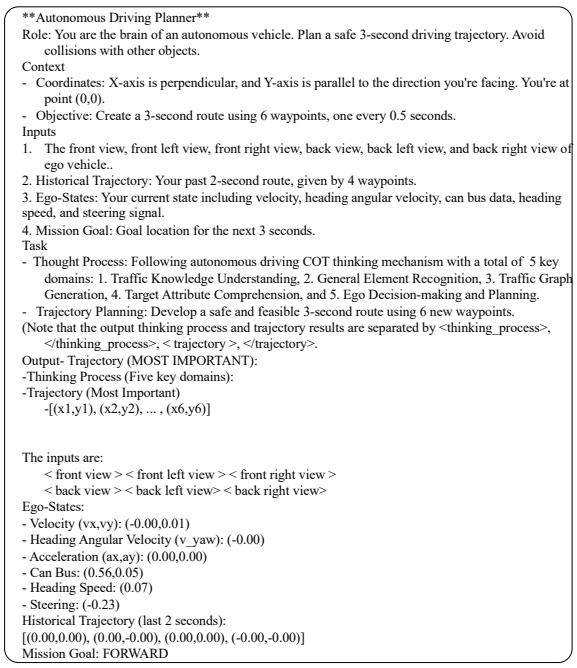


Figure 27. Examples of Trajectory.

Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Conference on Computer Vision and Pattern Recognition*, 2024. 1

[6] Xinpeng Ding, Jianhua Han, Hang Xu, Xiaodan Liang, Wei Zhang, and Xiaomeng Li. Holistic autonomous driving understanding by bird's-eye-view injected multi-modal large models. In *Conference on Computer Vision and Pattern Recognition*, 2024. 1

[7] Yunfei Guo, Fei Yin, Xiao-hui Li, Xudong Yan, Tao Xue, Shuqi Mei, and Cheng-Lin Liu. Visual traffic knowledge graph generation from scene images. In *International Conference on Computer Vision*, pages 21604–21613, 2023. 1

[8] Jinkyu Kim, Teruhisa Misu, Yi-Ting Chen, Ashish Tawari, and John Canny. Grounding human-to-vehicle advice for self-driving vehicles. In *Conference on Computer Vision and Pattern Recognition*, 2019. 1

[9] Yuhang Lu, Yichen Yao, Jiadong Tu, Jiangnan Shao, Yuexin Ma, and Xinge Zhu. Can lvlms obtain a driver's license? a benchmark towards reliable agi for autonomous driving. *arXiv preprint arXiv:2409.02914*, 2024. 1

[10] Yingzi Ma, Yulong Cao, Jiachen Sun, Marco Pavone, and Chaowei Xiao. Dolphins: Multimodal language model for driving. In *European Conference on Computer Vision*, 2024. 1, 2

[11] Srikanth Malla, Chiho Choi, Isht Dwivedi, Joon Hee Choi, and Jiachen Li. Drama: Joint risk localization and captioning in driving. In *Winter Conference on Applications of Computer Vision*, 2023. 1

[12] Jiageng Mao, Yuxi Qian, Junjie Ye, Hang Zhao, and Yue Wang. Gpt-driver: Learning to drive with gpt. *arXiv preprint arXiv:2310.01415*, 2023.

[13] Ana-Maria Marcu, Long Chen, Jan Hünermann, Alice Karnsund, Benoit Hanotte, Prajwal Chidananda, Saurabh Nair, Vijay Badrinarayanan, Alex Kendall, Jamie Shotton, et al. Lingoqa: Visual question answering for autonomous driving. In *European Conference on Computer Vision*, 2024. 1

[14] Chirag Parikh, Rohit Saluja, CV Jawahar, and Ravi Kiran Sarvadevabhatla. Idd-x: A multi-view dataset for ego-relative important object localization and explanation in dense and unstructured traffic. In *International Conference on Robotics and Automation*, 2024. 1

[15] Chonghao Sima, Katrin Renz, Kashyap Chitta, Li Chen, Hanxue Zhang, Chengen Xie, Ping Luo, Andreas Geiger, and Hongyang Li. Drivelm: Driving with graph visual question answering. In *European Conference on Computer Vision*, 2024. 1

[16] Huijie Wang, Tianyu Li, Yang Li, Li Chen, Chonghao Sima, Zhenbo Liu, Bangjun Wang, Peijin Jia, Yuting Wang, Shengyin Jiang, et al. Openlane-v2: A topology reasoning benchmark for unified 3d hd mapping. *Advances in Neural Information Processing Systems*, 36:18873–18884, 2023.

[17] Zhenhua Xu, Yujia Zhang, Enze Xie, Zhen Zhao, Yong Guo, Kwan-Yee K Wong, Zhenguo Li, and Hengshuang Zhao. Drivegpt4: Interpretable end-to-end autonomous driving via large language model. *IEEE Robotics and Automation Letters*, 2024. 1