Are VLMs Ready for Autonomous Driving? An Empirical Study from the Reliability, Data and Metric Perspectives

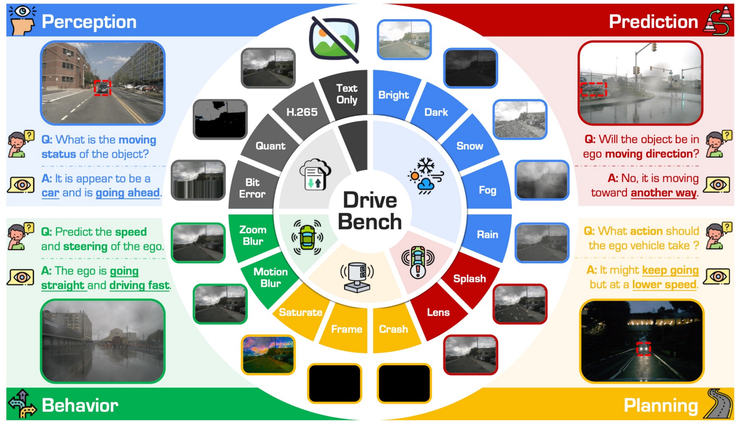
Shaoyuan Xie[1], Lingdong Kong[2,3], Yuhao Dong[2,4], Chonghao Sima[2,5], Wenwei Zhang[2], Qi Alfred Chen[1], Ziwei Liu[4], Liang Pan[2]

ICCV OCT 19-23, 2025 HONOLULU HAWAII

Motivation & Contribution

Benchmark Overview

- **DriveBench** is a benchmark designed to **reveal VLMs' limitations in autonomous driving**, It includes **19,200** frames and **20,498** QA pairs under **17** settings (**clean**, **corrupted**, and **text-only**), cover 4 driving tasks (perception, prediction, planning, and behavior).



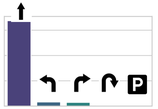
Motivation & Observation

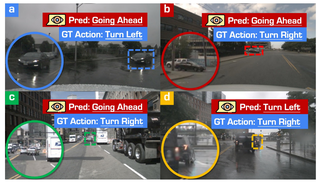
- **VLMs** are increasingly applied to different autonomous driving tasks for their **common-sense reasoning**, offering potential to address rare **corner cases** beyond data-driven coverage.
- However, their susceptibility to **hallucination** raises serious **safety concerns**, especially in safety-critical driving scenario.
- **DriveBench** tackles a critical yet unexplored assumption: ***"Can existing VLMs provide reliable, visually-grounded explanations?"***

1. UCI University of California, Irvine
2. 上海人工智能实验室 Shanghai Artificial Intelligence Laboratory OpenMMLab
3. NUS National University of Singapore
4. NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE MMLab @NTU
5. 香港大學 THE UNIVERSITY OF HONG KONG

Benchmark Construction

- **Dataset imbalance** is observed across multiple "driving with language" datasets, e.g., DriveLM, BDD-X. We resample the data in our benchmark construction to **make it more balanced**.





- We developed the main **DriveBench** based on DriveLM and remove data depends on **temporal information**, given the context window limitation.
- We filter the dataset based on GPT-4o results, removing "unreasonable" data for better testing existing VLMs.

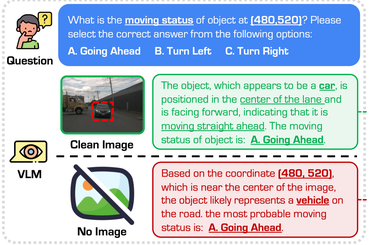
Reliability Check & Evaluation

Hallucination with Visual Degradation

- **VLMs fabricate answers** under **visual degradations** but show no GPT score degradation under corruption or text-only input.

Tab. GPT score with clean; corrupted; and no visual (Text-Only) input

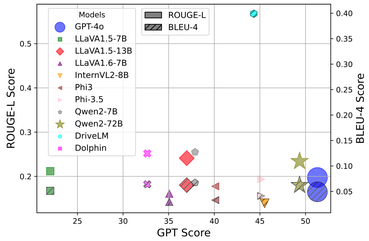
Method	Size	Type	Perception			Prediction			Planning			Behavior		
			Clean	Corr.	T.O.	Clean	Corr.	T.O.	Clean	Corr.	T.O.	Clean	Corr.	T.O.
Human	-	-	47.67	38.32	-	-	-	-	-	-	-	69.51	54.09	-
GPT-4o [2]	-	Commercial	35.37	35.25	36.48	51.30	49.94	49.05	75.75	75.36	73.21	45.40	44.33	50.03
LLaVA-1.5 [47]	7 B	Open	23.22	22.95	22.31	22.02	17.54	14.64	29.15	31.51	32.45	13.60	13.62	14.91
LLaVA-1.5 [47]	13 B	Open	23.35	23.37	22.37	36.98	37.78	23.98	34.26	34.99	38.85	32.99	32.43	32.79
LLaVA-NeXT [48]	7 B	Open	24.15	19.62	13.86	35.07	35.89	28.36	45.27	44.36	27.58	48.16	39.44	11.92
InternVL2 [12]	8 B	Open	**32.36**	32.68	33.60	45.52	37.93	48.89	53.27	55.25	34.56	**54.58**	40.78	20.14
Phi-3 [1]	4.2 B	Open	22.88	23.93	28.26	40.11	37.27	22.61	60.03	61.31	46.88	45.20	44.57	28.22
Phi-3.5 [1]	4.2 B	Open	27.52	27.51	28.26	45.13	38.21	4.92	31.91	28.36	46.30	37.89	49.13	39.16
Oryx [51]	7 B	Open	17.02	15.97	18.47	48.13	46.63	12.77	53.57	55.76	48.26	33.92	33.81	23.94
Qwen2-VL [71]	7 B	Open	28.99	27.85	35.16	37.89	39.55	37.77	57.04	54.78	41.66	49.07	47.68	54.48
Qwen2-VL [71]	72 B	Open	30.13	26.92	17.70	**49.35**	43.49	5.57	**61.30**	63.07	53.35	51.26	49.78	39.46



- We observe that existing driving VLMs can generate answer base on **text information** and **general knowledge** learned in training.
- This raises concerns about their **reliability** and **trustworthiness**, as such behaviors are often difficult to detect using existing datasets and evaluation metrics.

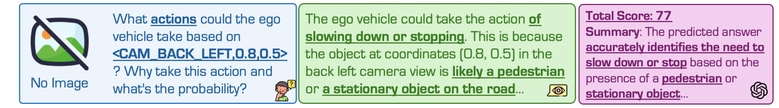
Benchmark Study & Experiments

Comparisons of Evaluation Metrics



- **Language metrics** (e.g., Rouge-L) fail to reflect **semantic similarity** and often show uniformly scores.
- **GPT scoring** is more discriminative, aligning with model performance the **standard benchmarks**.
- **Fine-tuned models** tend to "learn" the **answer template**, leading to a unwanted, misleading sense of performance improvement.

LLM-as-a-Judge



- GPT score is more **discriminative** with **more information** provided.
- Even with these the information, GPT score can also fail to capture nuance when the model is hallucination under **text-only** scenarios.

Robust Agentic Utilization (RAU)

Corruption Awareness

- VLMs tends to have the **corruption awareness** to **correctly** understand the current visual corruption type.
- Most VLMs acknowledge cannot answer questions when **explicitly include corruption type** in prompt.

Tab. Robustness improvement

Method	Input	NDS↑	mAP↑	mCE↓	mRR↑
DETR3D [73]	Clean	43.41	34.94	-	-
DETR3D [73]	Corrup.	30.76	19.26	1.22	0.71
$\text{DETR3D}_{\text{RAU}}$ [73]	Corrup.	34.12	22.72	**1.16**	**0.79**
BEVFormer [41]	Clean	51.71	41.63	-	-
BEVFormer [41]	Corrup.	30.64	20.13	1.23	0.59
$\text{BEVFormer}_{\text{RAU}}$ [41]	Corrup.	35.44	25.07	**1.14**	**0.68**

Tab. Corruption accuracy

Method						Avg
GPT-4o [2]	57.20	29.25	44.25	34.25	36.83	40.36
LLaVA-1.5_{7B} [47]	69.70	26.50	18.83	71.25	10.17	39.29
LLaVA-1.5_{13B} [47]	61.60	15.50	24.08	**79.75**	15.50	39.29
LLaVA-NeXT [48]	69.70	48.50	21.83	66.00	11.83	43.57
InternVL2 [12]	59.90	**50.75**	29.92	68.25	30.00	47.76
Phi-3 [1]	40.00	25.00	16.83	31.25	27.67	28.15
Phi-3.5 [1]	60.60	21.25	25.58	33.00	39.67	36.02
Oryx [51]	53.20	45.00	50.50	72.50	39.67	**52.17**
Qwen2-VL_{7B} [71]	**76.70**	37.50	22.83	57.00	35.83	45.97
Qwen2-VL_{72B} [71]	59.80	45.50	**52.25**	58.25	**44.83**	52.13
DriveLM [63]	21.20	**21.25**	9.00	**22.25**	17.50	18.24
Dolphins [52]	**54.30**	3.00	**9.42**	9.25	**21.50**	19.49

- **RAU leverages awareness of corruption** of driving VLMs to improve performance of some **downstream tasks** on corrupted visual input (e.g., 3D object detection).