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

Shaoyuan Xie[†] Lingdong Kong[‡,◊,∗] Yuhao Dong[‡,§] Chonghao Sima[‡,▽]
Wenwei Zhang[‡] Qi Alfred Chen[†] Ziwei Liu[§] Liang Pan[‡,✉]

[†]University of California, Irvine [‡]Shanghai AI Laboratory [◊]National University of Singapore
[§]S-Lab, Nanyang Technological University [▽]The University of Hong Kong

Code & Dataset: *https://drive-bench.github.io*

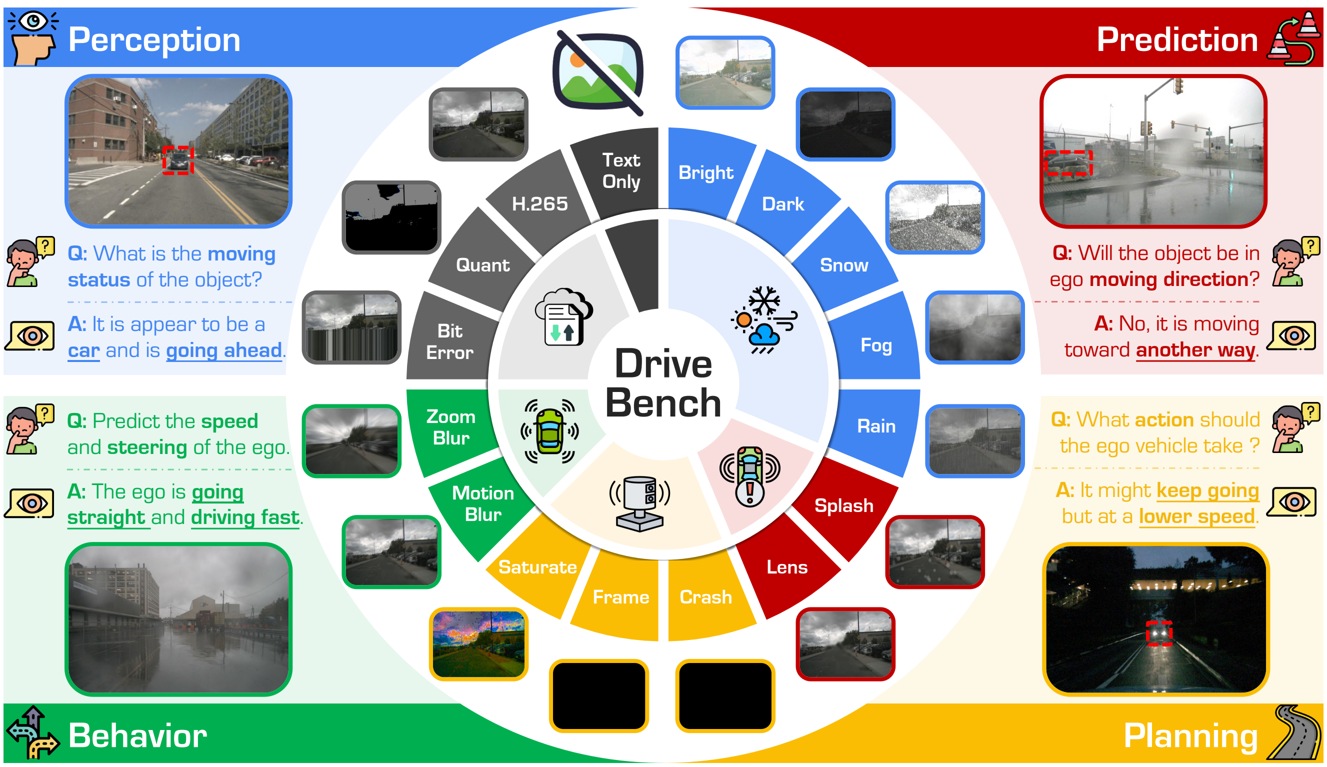

Figure 1. **Overview of DriveBench**. Our benchmark evaluates the reliability and visual grounding of Vision-Language Models (VLMs) in autonomous driving across four mainstream driving tasks – perception, prediction, planning, and behavior – under a diverse spectrum of **17 settings** (clean, corrupted, and text-only inputs). It includes **19,200 frames** and **20,498 QA pairs** spanning three question types: multiple-choice, open-ended, and visual grounding. By addressing diverse tasks and conditions, we aim to reveal VLMs' limitations and promote reliable, interpretable autonomous driving.

Abstract

Recent advancements in Vision-Language Models (VLMs) have fueled interest in autonomous driving applications, particularly for interpretable decision-making. However, the assumption that VLMs provide visually grounded and reliable driving explanations remains unexamined. To address this, we introduce DriveBench, a benchmark evaluating 12 VLMs across 17 settings, covering 19,200 images, 20,498 QA pairs, and four key driving tasks. Our findings reveal that existing VLMs often generate plausible responses from general knowledge or textual cues rather than true visual grounding, especially under degraded or missing visual inputs. This behavior, concealed by dataset imbalances and insufficient evaluation metrics, poses significant risks in safety-critical scenarios like autonomous driving. We further observe that VLMs possess inherent corruption-awareness but only explicitly acknowledge these

(∗) Project lead. (✉) Corresponding author.

issues when directly prompted. Given the challenges and inspired by the inherent corruption awareness, we propose Robust Agentic Utilization (RAU), leveraging VLMs' corruption awareness and agentic planning with external tools to enhance perception reliability for a diverse set of downstream tasks. Our study challenges existing evaluation paradigms and provides a road map toward more robust and interpretable autonomous driving systems.

1. Introduction

With recent advancements in Vision-Language Models (VLMs) [1, 2, 5, 12, 13, 46–48, 51, 71], there has been increasing research interest in applying VLMs to autonomous driving applications [20, 21, 26, 30, 43, 52, 53, 62, 63, 66, 72, 74, 77, 80, 84]. Recent research explores both integrating VLMs into end-to-end driving frameworks [20, 33, 57, 66, 72, 79], and extending VLMs into Vision-Language-Action (VLA) models that directly generate control commands [11, 22, 26, 30, 31, 62, 63, 77, 80, 86, 87]. This integration aims to leverage the common-sense reasoning capabilities of VLMs, learned from internet-scale knowledge, to improve the transparency and reliability of autonomous driving systems, especially in handling corner cases [82].

However, previous studies highlight significant limitations in evaluating end-to-end autonomous driving models in open-loop settings [42]. Instead of focusing on trajectory prediction with potentially unreliable open-loop end-to-end VLMs [33, 55, 63, 80], we address another fundamental – yet underexplored – question that has been widely assumed [55, 62, 66, 82]: *"Are existing VLMs capable of providing reliable explanations grounded on visual cues for driving?"*

To investigate, we examine whether driving decisions generated by VLMs are genuinely grounded in sensory information from the physical environment or reflect general knowledge and fabricated responses from textual cues.

VLM Reliability. To tackle the fundamental question above, we examine the model's reliability through an out-of-distribution (OoD) robustness lens. For this purpose, we introduce DriveBench, a benchmark encompassing four mainstream driving tasks and **15** types of data corruptions, including **19,200** images and **20,498** QA pairs tailored to real-world autonomous driving scenarios shown in Fig. 1. To assess robustness under extreme conditions, we push visual degradation to its extreme by using text-only prompts. Surprisingly, VLMs demonstrate comparable performance to their outputs under "clean" visual inputs, even when no visual cues are available (as illustrated in Fig. 2). This observation starkly contrasts with human drivers, who would struggle under such degraded conditions. A closer analysis reveals that this apparent "resilience" is often a byproduct of imbalanced datasets, suboptimal evaluation protocols, and model bias rather than actual model robustness.

Datasets. We perform an in-depth analysis of existing

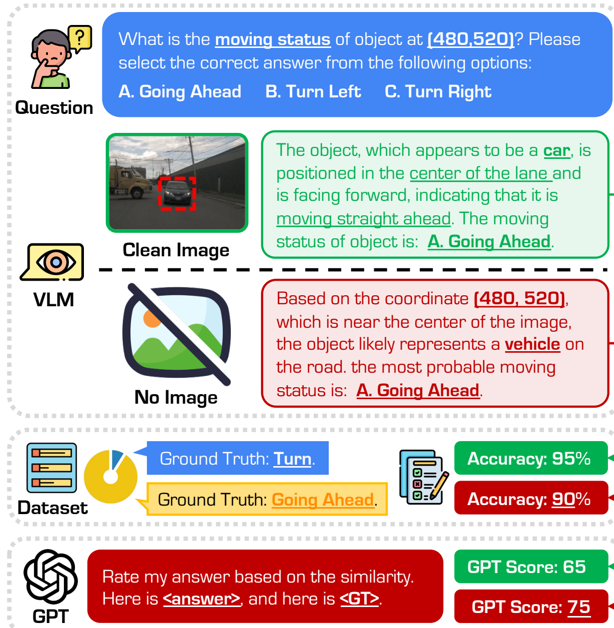

Figure 2. **Do VLMs provide reliable explanations based on visual cues in driving?** We study this from perspectives on reliability, data, and metrics. We find VLMs can fabricate quality answers to driving questions when visual information is absent. The fabricated answers can bypass current metrics, even GPT scores, due to imbalance, lack of a context dataset, and problematic evaluation protocols. Our observations challenge the passive assumption that VLMs are more reliable than task-specific models in driving decisions [28] because of visual-grounded interpretable responses.

"Driving with Language" benchmarks [9, 34, 59, 63, 78] and identify critical shortcomings, particularly concerning dataset imbalance. Many of these benchmarks, built on popular driving datasets such as nuScenes [59], BDD [85], and Waymo Open [65], inherit limitations from their original designs [42]. For instance, imbalanced data distributions skew evaluations, enabling overly simplistic answers such as *"Going Ahead"* to achieve over 90% accuracy for motion-related queries. Furthermore, some cases create challenges even for human annotators. Consequently, these benchmarks exhibit inherent biases and persistent negative samples, which diminish the interpretability and reliability of the evaluation and impair the model fine-tuned on them.

Metrics. We also revisit existing metric designs critically. Language interactions in driving applications are often assessed using traditional pattern-matching metrics such as ROUGE [44], BLEU [58], and CIDEr [69], which were originally developed for summarization and translation tasks. However, as noted in [3, 4, 18, 67], these metrics face significant limitations in evaluating nuanced language-based driving decisions. We also find that even GPT-based evaluators [10, 24, 49, 63] provide distinct scores given different prompts. These constraints underscore the urgent need for metrics that effectively capture reasoning, contextual understanding, and safety-critical aspects.

Through a series of comprehensive experiments, we derive several key insights from our analysis, spanning **17 settings** (*i.e.*, clean, text-only, and various corrupted inputs), **12 VLMs** (including both open-sourced and commercial models), **5 tasks** (perception, prediction, planning, behavior, and corruption identification), and **3 evaluation metrics** (accuracy scores, traditional language metrics [44, 58], and GPT scores). These findings shed light on the current challenges in integrating VLMs into driving scenarios:

1 **Fabricated responses under degradation:** VLMs often produce plausible yet fabricated responses under *degraded visual conditions*, including scenarios where no visual cues exist. This raises concerns about their reliability and trustworthiness, as such behaviors are difficult to detect using existing datasets and evaluation protocols.

2 **Awareness of visual corruptions:** While VLMs exhibit certain awareness of visual corruptions, they only acknowledge these issues when *directly prompted*. This highlights their limitations in assessing the reliability of inputs and providing scenario-specific, safety-focused responses.

3 **Impact of dataset biases:** Highly biased datasets and suboptimal evaluation protocols can create misleading impressions. In many cases, VLMs rely on general knowledge rather than actual visual cues to generate responses, which can unexpectedly achieve high scores with existing metrics.

4 **Need for tailored metrics:** Existing metrics, including language-based [44, 58] and GPT scores [10, 63], fail to capture the nuanced requirements of driving tasks. There is an urgent need for the development of specialized metrics that account for reasoning, contextual understanding, and safety-critical aspects to evaluate VLMs more effectively.

Our findings through DriveBench highlight the need for improved datasets, evaluation protocols, and more reliable VLMs. Motivated by these insights, we further propose *Robust Agentic Utilization* (RAU), leveraging VLM agents for enhanced perception in autonomous driving. RAU explores the potential of VLMs' corruption awareness and agentic planning with external tools to improve perception reliability, paving the way for more robust autonomous systems.

2. Related Work

Driving with Language. VLMs [1, 5, 46–48, 71] have demonstrated remarkable human-level reasoning and understanding across diverse domains [7, 11, 14, 16, 27, 45, 50, 64, 66, 79, 81, 88]. This capability has raised the prospect of utilizing VLMs to manage complex and unpredictable scenarios in autonomous driving [82]. Additionally, the language-based interaction that VLMs offer can help mitigate the black-box nature of deep neural networks by providing explanatory feedback that accompanies their decisions. Driven by these advantages, a growing body of research has begun building benchmarks of VLMs in autonomous driving [34, 54, 59, 63, 75, 78]. However, despite these advancements, the robustness and reliability of VLMs in complex, real-world autonomous driving tasks remain largely untested, especially given that reliable performance across diverse driving situations is a fundamental requirement for their application in autonomous driving.

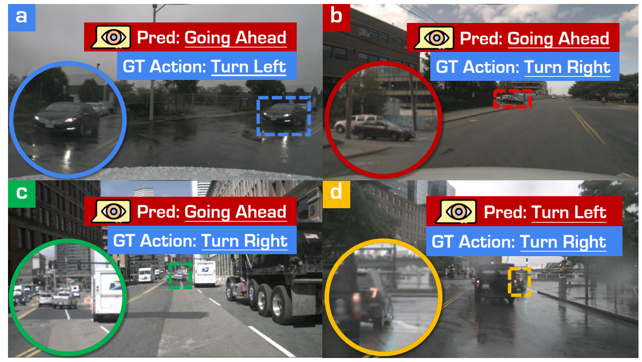

Figure 3. **Challenging cases excluded from DriveBench.** The results are from GPT4o [2]. **(a)**: A black sedan is turning left, indicated by the turn lights. **(b)**: A black sedan is turning right. The model predicts both *Going Ahead*. The examples show challenging cases for *Turn* choice, where the visual cues are subtle or rely on temporal context for correct predictions. **(c)** and **(d)** are both *Turning Right*, but the model fails to locate the objects due to the existence of overlapping or occlusion.

VLM Reliability. Deep neural networks have historically struggled with out-of-distribution (OoD) data, a limitation of particular concern in autonomous driving, where failing to handle rare or unexpected scenarios could result in severe consequences [35, 36, 76]. While existing research attempted to explore VLM hallucinations and trustworthiness [32, 40, 68, 70], it has not yet been rigorously examined within the context of driving applications. Autonomous driving raises new challenges to evaluate the reliability of VLMs where language-based driving decisions are naturally linked to physical and context-specific real-world scenarios. In this work, we provide a systematic evaluation of the reliability of current VLMs under conditions of visual corruption, identifying potential limitations that impact their applicability in real-world driving.

3. DriveBench: Driving with VLMs

In this section, we detail the construction of our benchmark designed to assess the reliability of VLMs within the domain of autonomous driving. The comparison between our dataset and related benchmarks is presented in Tab. 1.

3.1. Datasets

We construct our benchmark with representative driving with language datasets [63]. We choose DriveLM [63] as it is acknowledged as one of the most representative datasets for driving with languages [17, 56]. The dataset spans five

Table 1. **Comparisons among evaluation benchmarks** for driving. "**Per.**", "**Pre.**", "**Beh.**", "**Pla.**", "**Rob.**" refer to the Perception, Prediction, Behavior, Planning, and Robustness tasks, respectively. GPT_{ctx} represents GPT scores augmented with context information.

Benchmark	**Per.**	**Pre.**	**Beh.**	**Pla.**	**Rob.**	**# Frames** (Test Data)	**# QA Pairs** (Test Data)	**Logic**	**Evaluation Metrics** Acc	Language	F1	GPT	GPT_{ctx}
BDD-X [34]	✓	✗	✗	✗	✗	-	-	None	No	Yes	No	No	No
BDD-OIA [78]	✓	✗	✓	✗	✗	-	-	None	No	No	Yes	No	No
nuScenes-QA [59]	✓	✗	✗	✗	✗	36,114	83,337	None	Yes	No	No	No	No
Talk2Car [15]	✓	✗	✗	✓	✗	$\sim$ 1.8K	2,447	None	Yes	No	No	No	No
nuPrompt [75]	✓	✗	✗	✗	✗	$\sim$ 36K	$\sim$ 6K	None	Yes	No	No	No	No
DRAMA [54]	✓	✗	✗	✓	✗	-	$\sim$ 14K	Chain	No	Yes	No	No	No
Rank2Tel [61]	✓	✗	✗	✓	✗	-	-	Chain	Yes	Yes	No	No	No
DirveMLLM [25]	✓	✗	✗	✗	✗	880	-	None	Yes	No	No	No	No
DriveVLM [66]	✓	✗	✓	✓	✗	-	-	None	No	No	No	No	Yes
DriveLM [63]	✓	✓	✓	✓	✗	4,794	15,480	Graph	No	Yes	Yes	No	No
DriveBench	✓	✓	✓	✓	✓	**19,200**	**20,498**	**Graph**	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**

tasks, including perception, prediction, planning, behavior, and control. For each task, different sets of questions are applied, such as multiple-choice questions (MCQs), and visual question answering (VQA). For clarity, we will use {*Task*}-{*Question Type*} to specify the data in the rest of the paper (*e.g.*, perception-MCQs).

Distribution Bias. Through detailed examination, we identify a significant distribution bias in the dataset, which is naturally inherited from the nuScenes dataset [6, 42]. Specifically, in behavior-MCQs that inquire about the future movement of the ego vehicles, approximately 78.6% of responses are labeled as *"Going Ahead"*, which severely impair the evaluation and induce bias towards the fine-tuned model as studied in Appendix A.1. To address this imbalance, in DriveBench, we carefully re-sampled the data to create a more balanced distribution among different options. The detailed distribution can be found in Appendix B.1. We also investigate BDD-X [34, 85] dataset and find that bias commonly exists in current driving with language benchmarks, detailed analysis can be found in Appendix A.2.

Challenging Cases. Furthermore, we evaluate GPT-4o [2] and analyze its failure cases, as illustrated in Fig. 3. We find cases such as *"Turn Left"* or *"Turn Right"* are factually correct but involve (a) long temporal context; (b) subtle indicators (*e.g.*, the turn signal); (c) overlapping, and (d) occlusion, which is confusing even for human at first glance. It is more concerning given the input length constraint of image resolutions and temporal lengths of existing VLMs. Therefore, we eliminate these outlier instances to prevent such samples from obscuring our findings and focus on analyzing the average cases. Due to space limits, more details can be found in the case study in Appendix E.4.

3.2. Driving Tasks

Our DriveBench covers four mainstream driving tasks, including **perception**, **prediction**, **planning**, and **behavior**, examples are shown in Fig. 1. The definition and distribution of each task can be found in Appendix B.3.

3.3. Corruption Data

We craft a total of **15** visual corruption types (*cf*. Fig. 1), spanning across **weather conditions** ([1]`Brightness`, [2]`Dark`, [3]`Fog`, [4]`Snow`, and [5]`Rain`), **external disturbances** ([6]`Water Splash` and [7]`Lens Obstacle`), **sensor failures** ([8]`Camera Crash`, [9]`Frame Lost`, and [10]`Saturate`), **motion blurs** ([11]`Motion Blur` and [12]`Zoom Blur`), and **data transmission errors** ([13]`Bit Error`, [14]`Color Quant`, and [15]`H.265 Compression`). We encompass a range of potential OoD scenarios the vehicles might encounter [36, 37, 76]. From a reliability perspective, these corruptions are the key to our evaluation and insights into VLMs' visual-grounded driving capabilities. For more detailed corruption definitions and the generation process, please refer to Appendix B.2.

3.4. Vision-Language Models (VLMs)

To encompass the full scope of existing advanced VLMs, the current version of DriveBench evaluates a diverse set of **12** popular VLMs, including both commercial and open-source models, as well as models fine-tuned specifically for autonomous driving applications [52, 63]. This selection reflects the latest developments in state-of-the-art VLMs for driving. To ensure consistency, we apply a standardized system prompt across all models (further prompt details are provided in the Appendix C.2). The prompt explicitly instructs the VLMs to generate auxiliary explanations, enabling GPT-based evaluation of single-answer MCQs.

3.5. Evaluation Metrics

We consider a comprehensive set of metrics, including Accuracy, BLEU [58], ROUGE-L [44], and GPT scores [10, 63]. For MCQs, we utilize both accuracy, as the most direct measure, and GPT scores to capture nuances in the explanatory quality beyond simple answer selection. For VQAs, we choose BLEU, ROUGE-L, and GPT scores. We further improve the GPT evaluation in [63] by providing detailed rubrics, scenario-based context, denoted as GPT_{cxt}.

Table 2. **Evaluations of VLMs across different driving tasks** (perception, prediction, planning, and behavior). "Clean" represents clean image inputs. "Corr." represents corruption image inputs, averaged across fifteen corruptions. "T.O." represents text-only evaluation. For humans, only perception-MCQ and behavior-MCQ are evaluated. The evaluations are based on GPT_{cxt} scores, where we tailored detailed rubrics for each task and question type. We highlight scores higher than the corresponding clean performance under corruptions.

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	<u>60.03</u>	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	<u>48.13</u>	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	<u>30.13</u>	26.92	17.70	**49.35**	43.49	5.57	**61.30**	63.07	53.35	<u>51.26</u>	49.78	39.46
DriveLM [63]	7 B	Specialist	**16.85**	16.00	8.75	**44.33**	39.71	4.70	**68.71**	67.60	65.24	**42.78**	40.37	27.83
Dolphins [52]	7 B	Specialist	9.59	10.84	11.01	32.66	29.88	39.98	52.91	53.77	60.98	8.81	8.25	11.92

4. Experiments & Analyses

We conduct extensive benchmark experiments and analyses in DriveBench, with detailed discussions leading to our observations and conclusions by step.

4.1. Experimental Setups

Models. We set the temperature to 0.2 and `top-p` to 0.2, with a maximum output token limit of 512. For DriveLM-Agent [63], we adhere to the configurations outlined in [17]. Specifically, we utilize LLaMA-Adapter-V2 [23] as the base model, fine-tuned on the DriveLM-nuScenes dataset. The fine-tuning process is conducted on A800 GPUs with a batch size of 4, over 4 epochs. For other open-source models, we download the official model weight from HuggingFace and inference using the `vLLM` [38] framework. More details about the used model configuration can be found in Appendix C.1. For GPT-4o, we query the official APIs from OpenAI with the same configuration mentioned above. The model is provided with single-frame images by default. We also show the generality of our observation under multi-frame temporal input in Appendix E.2. Additionally, we provide the single-view image if only that view is required.

Metrics. For GPT score evaluation, we employ GPT-3.5-turbo. To better capture nuances between responses, we prompt the model with detailed rubrics that account for answer correctness, coherence, and the alignment of explanations with the final answer. Rubrics are designed for each specific task and question type to better reflect human-preferred responses. Detailed information on the GPT evaluation prompts and rubrics can be found in Appendix C.3.

4.2. Observations & Discussions

We mainly report GPT_{cxt} scores in the rest of the paper unless otherwise specified. Due to space limits, the complete results with different metrics are provided in Appendix D.

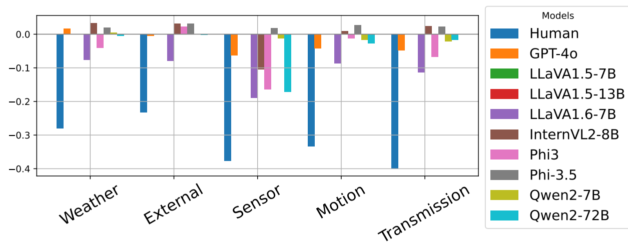

Figure 4. **Illustration of performance degradation**. After applying each corruption, we evaluate the perception-MCQs accuracy changes compared with clean inputs. We observe that human performance largely decreases while most VLMs remain unchanged.

4.2.1. Corruption Resilience

The primary results, evaluated using GPT_{cxt}, are summarized in Tab. 2. We observe that, even in the presence of corruption, the model performance remains largely unaffected, demonstrating "seemingly" resilience to such OoD scenarios. Specifically, a noticeable portion of VLMs maintains comparable performance to that with clean image inputs, even in open-ended VQAs. To understand the source of the resilience, we investigate whether it stems from the robustness of these VLMs, given their large-scale pre-training data [19], or if other factors contribute to this phenomenon.

Human Evaluations. To further validate that the applied corruptions indeed impact the driving scenario, we conduct a human evaluation. Specifically, we sub-sample the dataset and design a user interface to facilitate human performance assessment (more details in Appendix C.4). The accuracy degradation is shown in Fig. 4. Interestingly, we observe a significant accuracy drop for human participants under corrupted conditions, whereas most VLMs exhibit subtle performance variations across different corruption types.

Text-Only Prompts. Given the above results, we fur-

Table 3. **Comparisons of perception-MCQ and behavior-MCQ accuracy scores between "clean" and fully "black" (no image) inputs**. We observe a large portion of models have no clear performance degradation even when the visual information is absent, suggesting the driving VLMs response might mainly be based on general knowledge, instead of leveraging specific visual cues from sensors.

Task	Image	Human	GPT-4o [2]	LLaVA-NeXT [48]	LLaVA-1.5$_{13B}$ [47]	Phi-3 [1]	Phi-3.5 [1]	Qwen2-VL$_{7B}$ [71]	Qwen2-VL$_{72B}$ [71]
Perception	Clean	93.3	59.0	55.0	50.0	54.5	56.5	59.0	60.0
	No Image	-	59.5 $\uparrow$ 0.5	34.5 $\downarrow$ 20.5	50.0 $\downarrow$ 0.0	17.5 $\downarrow$ 37.0	58.5 $\uparrow$ 2.0	56.5 $\downarrow$ 2.5	23.5 $\downarrow$ 36.5
Behavior	Clean	69.5	25.5	33.5	32.5	26.5	36.5	30.0	23.0
	No Image	-	24.0 $\downarrow$ 1.5	24.0 $\downarrow$ 9.5	33.0 $\uparrow$ 0.5	30.0 $\uparrow$ 3.5	40.0 $\uparrow$ 3.5	23.0 $\downarrow$ 7.0	36.5 $\uparrow$ 13.5

Table 4. **Comparisons of perception-MCQ accuracy degradation after prompting VLMs with explicit corruption context**. We notice a clear trend of performance degradation after mentioning the corruption type in the question. The results suggest VLMs are aware of the current corruption and acknowledge they can not respond due to the degraded visual information when explicitly prompted.

Method	Bright	Dark	Snow	Fog	Rain	Lens	Water	Cam	Frame	Saturate	Motion	Zoom	Bit	Quant	H.265
GPT-4o	−8.69	−12.98	−8.25	−9.00	−6.00	−3.81	−5.82	−12.94	−10.99	−8.52	−6.98	0.57	−8.22	−4.79	−14.30
LLaVA-1.5$_{7B}$	0.26	1.04	0.25	0.00	0.00	1.40	2.60	−2.79	−8.97	0.51	−0.52	2.57	2.22	−1.32	−2.66
LLaVA-1.5$_{13B}$	0.26	1.04	0.25	0.00	0.00	1.96	2.60	−1.27	−0.26	0.51	1.04	2.57	2.22	−0.26	−2.07
LLaVA-NeXT	−5.83	−20.63	−31.95	−14.00	−18.50	−31.39	−36.97	−6.13	−18.29	−17.67	−24.85	−33.29	−19.50	5.89	−21.19
InternVL$_{8B}$	−7.24	−8.92	−10.74	−9.50	−7.50	−7.54	−6.24	−17.51	−0.23	−2.46	−2.35	−7.00	−6.67	−7.71	−4.65
Phi-3.5	−9.78	−7.48	−7.75	−9.00	−8.50	−8.60	−7.48	−16.37	−9.31	−9.50	−8.48	−8.07	−6.94	−11.29	−11.16
Phi-3	-4.22	8.67	0.75	−5.00	−10.00	−11.31	−33.22	3.03	8.29	−8.51	−5.42	3.57	17.89	−18.81	-13.12
Qwen2-VL$_{7B}$	−9.74	−7.96	−9.75	−9.50	−9.00	−5.93	−6.98	−20.94	−29.85	−8.49	−8.46	−3.00	−5.06	−9.38	−11.07
Qwen2-VL$_{72B}$	−6.70	−8.96	−8.25	−9.50	−11.00	−8.04	−6.90	7.19	11.01	−10.51	−7.44	−2.93	−6.61	−9.29	−13.07

ther investigate the effects of extreme corruption by providing VLMs with fully black images, reducing the input to text-only prompts with no visual information. The results, shown in Tab. 2, reveal an intriguing pattern: GPT$_{\mathrm{cxt}}$ scores for text-only prompts are closely aligned with those obtained with clean image inputs. This trend persists across different tasks and models, suggesting that the seeming resilience is not solely due to the inherent robustness.

We also report the accuracy for the perception-MCQs, as shown in Tab. 3. Surprisingly, a significant portion of the models show minimal or no accuracy degradation, even in the complete absence of visual cues. Upon further examination, we observe that the "resilience" of VLMs under text-only conditions is likely influenced by the extensive general knowledge acquired during training. For instance, the models can "guess" the moving status of one surrounding object based on text cues referring to which camera it has been seen and the corresponding position in that image. An example is shown in Fig. 5. To justify the generality of the findings and exclude text cues, we also study the visual-based object prompt (*i.e.*, using a visualized bounding box to specify a certain object), detailed in Appendix E.1. In summary, these observations yield **two key insights**:

- VLMs are capable of producing plausible responses to driving-related questions based solely on general knowledge or text prompts. This capability is likely attributed to the extensive general knowledge and common-sense reasoning capabilities acquired during their training.
- The current evaluation protocols for assessing VLMs in autonomous driving reveal significant shortcomings. Even advanced evaluation methods, such as GPT score, fail to effectively reflect the reliability of driving VLMs based on specific real-world scenarios.

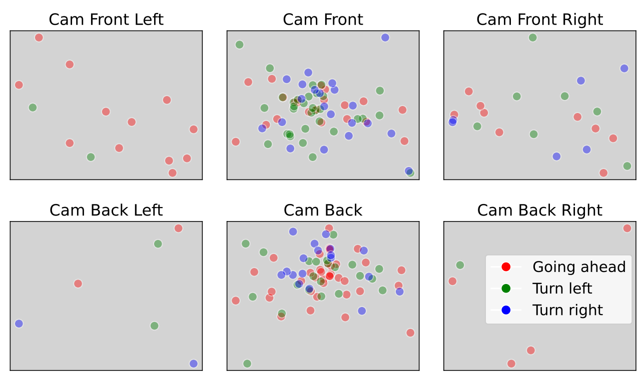

Figure 5. **Perception-MCQs answer spatial distribution** of Qwen2-VL$_{7B}$ [71] under **text-only prompts**. We visualize the MCQs prediction given the object's spatial position on different cameras. The model can potentially *"guess"* the answers without visual information by leveraging text cues. For example, *"what is the moving status of object at (480, 520) in front camera?"*. We also study the visual-based object prompt (*i.e.*, using a visualized bounding box to specify an object), detailed in Appendix E.1. More model case studies are included in Appendix Fig. G.4.

To investigate the first insight further, we pose the question: *"Are driving VLMs aware of the underlying corruptions in images when they fabricate their answers?"*

4.2.2. Corruption Awareness

We explore whether the fabricated "reasonable" answer of VLMs under corruption might stem from a lack of *awareness* regarding potential visual corruptions. To investigate this, we conduct two experiments: **E-1)** involves explicit corruption reference when prompting the model, *e.g.*, *"what are the important objects **in the snowy day**"*, and **E-2)** we directly ask the model to identify the current type of image corruption, *e.g.*, *"what is the current corruption"*.

Table 5. **Study on corruption awareness (robustness-MCQs).** We directly prompt VLMs to identify the type of corruption and average the accuracy score within each corruption type (defined in Sec. 3.3): weather conditions, external disturbances, sensor failures, motion blurs, and data transmission errors.

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

In E-1, we analyze changes in perception-MCQs accuracy. As shown in Tab. 4, the results demonstrate a notable trend of decreasing accuracy across various models and corruption types. Certain models exhibit substantial performance declines in the presence of corruption prompts; for example, LLaVA-NeXT$_{7B}$ [48] experiences an accuracy reduction of approximately 19.62%. A closer examination of model responses reveals increased uncertainty when the corruption context is included in the prompt. For instance, the model may respond with a statement such as *"based on the image, it is not possible to determine the moving status of the object..."*. These findings suggest that some models exhibit a degree of corruption awareness when explicitly prompted, recognizing potential unreliability in their responses under conditions of severe visual degradation.

Conversely, models such as LLaVA-1.5 [47] exhibit minimal performance changes even when corruption-specific prompts are provided. This observation, when combined with the previous findings, suggests two possible explanations: 1) these models may lack the capability to detect image corruption, or 2) while aware of the corruption, their responses remain dominated by general knowledge rather than current visual information, even in clean situations.

To investigate the first hypothesis, we conduct E-2, in which we explicitly prompt the VLMs to identify the type of visual corruption, which we call robustness-MCQs for naming consistency. The results in Tab. 5 indicate that LLaVA-1.5 [47] achieves competitive accuracy in identifying corruption types, particularly in weather and motion corruptions, suggesting it possesses corruption awareness.

To study the second hypothesis, we analyze the confusion matrix of responses from LLaVA-1.5 [47] in the perception-MCQs. Remarkably, the model consistently outputs *"Going Ahead"*, regardless of the actual visual context (visualized in Fig. G.4 in Appendix). This uniformity in answering indicates the model response is biased toward general knowledge rather than relying on current visual information. Therefore, combining the results with the findings in Sec. 4.2.1, we **conclude** below:

- VLMs tend to rely predominantly on common sense or text-based cues to generate responses under conditions of visual degradation, even though they are aware of it.

4.2.3. Fine-Tuned VLMs

In this section, we mainly focus on VLMs fine-tuned specifically on driving datasets, reflecting the growing body of research dedicated to this area [52, 63, 66]. Specifically, we select DriveLM [63] and Dolphin [52] as representative models for our analysis, as both are fine-tuned to enhance visual-grounded driving decision-making abilities.

The main results are summarized in Tab. 2. A key observation is that Dolphin [52], which is primarily fine-tuned on the BDD [85] dataset, demonstrates significant difficulty in answering questions from the nuScenes [59] dataset. Given the general capabilities of VLMs to address questions across diverse domains, this result is both surprising and concerning, highlighting the limited generalizability of driving-specific VLMs when exposed to datasets or question formats that differ from their fine-tuning conditions. Regarding DriveLM [63], we further investigate how the model benefits from in-distribution fine-tuning. We visualize the results from different metrics towards the same answer in Fig. 6. DriveLM [63], while surpassing other VLMs with large margins under ROUGE-L evaluation, still lags behind Qwen2-VL$_{72B}$ [71] and GPT-4o [2] in GPT evaluation. The observation indicates that the main improvement of in-distribution fine-tuning on the current small-scale driving dataset largely comes from the answering template. This analysis aims to elucidate the potential advantages and limitations of fine-tuning on a specific language-annotated driving dataset.

4.2.4. Metrics

Evaluating open-ended answers is still a challenging problem [8, 60, 83]. The problem is further escalated in driving, given that the safety of vehicle decisions is closely connected to a specific physical environment. To better understand the existing metrics' applicability in driving, We experiment with the same response under different evaluation metrics, including accuracy, language metrics, GPT score, and GPT$_{cxt}$ score. The results suggest that the same response evaluated under different metrics can vary significantly. Even using LLM-as-Judge with different prompts can lead to different results. We argue that existing metrics are far from enough to effectively reflect the reliability of driving VLMs. We provide full evaluation results in Appendix D. Due to space limits, additional analyses on the relationship between accuracy *vs*. GPT score, language metric *vs*. GPT score, and GPT score *vs*. GPT$_{cxt}$ score can be found in the Appendix E.3.

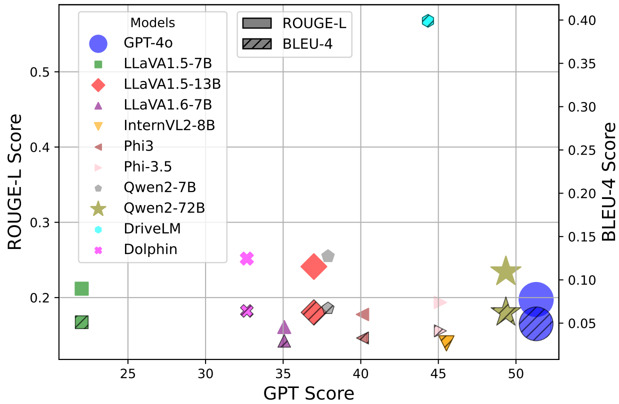

Figure 6. **Prediction-VQA evaluations using different metrics.** The language metrics, such as ROUGE-L [44] and BLEU-4 [58], exhibit high consistency, while $\mathrm{GPT}_{\mathrm{ctx}}$ scores demonstrate noticeable gaps. We also observe that fine-tuned process benefits DriveLM [23, 63] significantly in regulating its response format, thus leading to misleadingly high performance under language metrics.

5. Robust Agentic Utilization (RAU)

Given the observed drawbacks of existing benchmarks, metrics, and models, while inspired by the corruption awareness above, we explore how the inherent robustness awareness can be leveraged toward robust perception in autonomous driving. Specifically, we focus on developing *Robust Agentic Utilization* (RAU), applying VLMs as agents augmented with tools for robust perception.

Previous research shows the trade-off between OoD robustness and performance [76]. Meanwhile, the denoise-based approach is not extensible as separate training is needed given new corruption types [39]. Inspired by the corruption awareness of VLMs. We instead explore the use of VLMs as an agentic interface for robust perception.

5.1. Approach

Without losing generality, this paper focuses on the usage of RAU on one downstream task, camera-based 3D object detection [29, 41, 73], as it serves as the first component in full-stack autonomous driving pipelines. For the tools, we choose the denoise model [39] to restore the visual information. We train a denoise model for each of the corruptions and assemble them as tools. Then, we use VLMs as the planner to decide which one to use at run-time. This framework is extensible since a new denoiser can add flexibility and does not require re-training downstream models for robustness. Additionally, the environmental conditions in real-world autonomous driving do not change from frame to frame. Therefore, the inference cost for RAU is needed only when the environment changes. Furthermore, developing RAU is orthogonal to VLM and tool evolution: our framework can continuously benefit from the progress of VLMs and available model tools (*e.g.*, the denoise model).

Table 6. **RAU robustness evaluation**. mCE and mRR metrics are only applied to robustness evaluation. For mCE, we choose DETR3D [73] as the baseline. Detailed definition of metrics can be found in RoboBEV benchmark [76]. Equipped with RAU, we can improve the robustness of BEV detectors under corruption.

Method	Input	NDS↑	mAP↑	mCE↓	mRR↑
DETR3D [73]	Clean	43.41	34.94	-	-
DETR3D [73]	Corrup.	30.76	19.26	1.22	0.71
$\mathrm{DETR3D}_{\mathrm{RAU}}$ [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
$\mathrm{BEVFormer}_{\mathrm{RAU}}$ [41]	Corrup.	35.44	25.07	**1.14**	**0.68**

5.2. Setups

We evaluate the approach using camera-based 3D object detection model [41, 73] on RoboBEV benchmark [76]. The robustness evaluation is averaged across six different corruptions, including `Bright`, `Dark`, `Fog`, `Snow`, `Color Quant`, and `Motion Blur`. More details on the denoising model training and denoising qualitative results can be found in Appendix C.5. We use InternVL2 [12] as the agentic VLM without losing generality.

5.3. Results

Our RAU can largely improve the robustness under corruptions to downstream BEV detectors. Specifically, $\mathrm{BEVFormer}_{\mathrm{RAU}}$ and $\mathrm{DETR3D}_{\mathrm{RAU}}$ improve the NDS by 10.9% and 15.6%, respectively. The results can be potentially further boosted by improving the VLMs and the denoising model, which is out of the scope of this paper. Detailed results of RAU corruption identification accuracy and BEV detector performance for each corruption are presented in Appendix D.4. Besides 3D detection, the RAU can potentially be used for end-to-end driving [62, 63], or even used before the images are input to the VLMs themselves, which we leave as future work. We hope our initial efforts can inspire future works exploring for trustworthy integration of VLMs in autonomous driving.

6. Conclusion

This work identifies and addresses key challenges in deploying Vision-Language Models (VLMs) for autonomous driving, with an emphasis on their visual grounding reliability in complex real-world scenarios. Our findings reveal that VLMs frequently generate plausible yet unsupported responses when subjected to visual degradation, casting doubt on their reliability in critical decision-making tasks in autonomous driving. Furthermore, imbalanced datasets and suboptimal evaluation amplify these concerns, contributing to an overestimation of VLM reliabilities. Finally, we propose Robust Agentic Utilization (RAU) inspired by corruption awareness to improve perception reliability in autonomous driving under visual corruption.

Acknowledgments

This study is supported by the Shanghai Artificial Intelligence Laboratory.

This study is supported by the Ministry of Education, Singapore, under its MOE AcRF Tier 2 (MOE-T2EP20221-0012, MOE-T2EP20223-0002), and under the RIE2020 Industry Alignment Fund – Industry Collaboration Projects (IAF-ICP) Funding Initiative, as well as cash and in-kind contributions from the industry partner(s). This work is also supported by National Key Research and Development Program of China (2024YFE0210700).

Lingdong Kong is supported by the Apple Scholars in AI/ML Ph.D. Fellowship program.

Additionally, the authors would like to sincerely thank the Program Chairs, Area Chairs, and Reviewers for the time and effort devoted during the review process.

References

[1] Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Harkirat Behl, et al. Phi-3 Technical Report: A Highly Capable Language Model Locally on Your Phone. *arXiv preprint arXiv:2404.14219*, 2024.

[2] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. GPT-4 Technical Report. *arXiv preprint arXiv:2303.08774*, 2023.

[3] Mousumi Akter, Naman Bansal, and Shubhra Kanti Karmaker. Revisiting Automatic Evaluation of Extractive Summarization Task: Can We Do Better Than ROUGE? In *Findings of the Association for Computational Linguistics*, pages 1547–1560. Springer, 2022.

[4] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. SPICE: Semantic Propositional Image Caption Evaluation. In *European Conference on Computer Vision*, pages 382–398. Springer, 2016.

[5] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. Qwen Technical Report. *arXiv preprint arXiv:2309.16609*, 2023.

[6] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuScenes: A Multimodal Dataset for Autonomous Driving. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11621–11631, 2020.

[7] Boyuan Chen, Zhuo Xu, Sean Kirmani, Brain Ichter, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. SpatialVLM: Endowing Vision-Language Models with Spatial Reasoning Capabilities. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14455–14465, 2024.

[8] Guiming Hardy Chen, Shunian Chen, Ziche Liu, Feng Jiang, and Benyou Wang. Humans or LLMs as the Judge? A Study on Judgement Bias. In *Conference on Empirical Methods in Natural Language Processing*, pages 8301–8327, 2024.

[9] 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. In *IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 7817–7826, 2025.

[10] Long Chen, Oleg Sinavski, Jan Hünermann, Alice Karnsund, Andrew James Willmott, Danny Birch, Daniel Maund, and Jamie Shotton. Driving with LLMs: Fusing Object-Level Vector Modality for Explainable Autonomous Driving. In *IEEE International Conference on Robotics and Automation*, pages 14093–14100, 2024.

[11] Yuntao Chen, Yuqi Wang, and Zhaoxiang Zhang. DrivingGPT: Unifying Driving World Modeling and Planning with Multi-Modal Autoregressive Transformers. *arXiv preprint arXiv:2412.18607*, 2024.

[12] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. InternVL: Scaling Up Vision Foundation Models and Aligning for Generic Visual-Linguistic Tasks. *arXiv preprint arXiv:2312.14238*, 2023.

[13] Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. How Far Are We to GPT-4V? Closing the Gap to Commercial Multimodal Models with Open-Source Suites. *Science China Information Sciences*, 67(12): 220101, 2024.

[14] Can Cui, Yunsheng Ma, Xu Cao, Wenqian Ye, and Ziran Wang. Drive as You Speak: Enabling Human-Like Interaction with Large Language Models in Autonomous Vehicles. In *IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 902–909, 2024.

[15] Thierry Deruyttere, Simon Vandenhende, Dusan Grujicic, Luc Van Gool, and Marie-Francine Moens. Talk2Car: Taking Control of Your Self-Driving Car. In *Conference on Empirical Methods in Natural Language Processing*, pages 2088–2098, 2019.

[16] Yuhao Dong, Zuyan Liu, Hai-Long Sun, Jingkang Yang, Winston Hu, Yongming Rao, and Ziwei Liu. Insight-V: Exploring Long-Chain Visual Reasoning with Multimodal Large Language Models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9062–9072, 2025.

[17] DriveLM contributors. DriveLM: Driving with Graph Visual Question Answering. https://github.com/OpenDriveLab/DriveLM, 2023.

[18] Mikhail Evtikhiev, Egor Bogomolov, Yaroslav Sokolov, and Timofey Bryksin. Out of the BLEU: How Should We Assess Quality of the Code Generation Models? *Journal of Systems and Software*, 203:111741, 2023.

[19] Alex Fang, Gabriel Ilharco, Mitchell Wortsman, Yuhao Wan, Vaishaal Shankar, Achal Dave, and Ludwig Schmidt. Data Determines Distributional Robustness in Contrastive Language Image Pre-Training (CLIP). In *International Conference on Machine Learning*, pages 6216–6234. PMLR, 2022.

[20] Bowen Feng, Zhiting Mei, Baiang Li, Julian Ost, Roger Girgis, Anirudha Majumdar, and Felix Heide. VERDI: VLM-Embedded Reasoning for Autonomous Driving. *arXiv preprint arXiv:2505.15925*, 2025.

[21] Daocheng Fu, Xin Li, Licheng Wen, Min Dou, Pinlong Cai, Botian Shi, and Yu Qiao. Drive Like a Human: Rethinking Autonomous Driving with Large Language Models. In *IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 910–919, 2024.

[22] Haoyu Fu, Diankun Zhang, Zongchuang Zhao, Jianfeng Cui, Dingkang Liang, Chong Zhang, Dingyuan Zhang, Hongwei Xie, Bing Wang, and Xiang Bai. Orion: A Holistic End-to-End Autonomous Driving Framework by Vision-Language Instructed Action Generation. *arXiv preprint arXiv:2503.19755*, 2025.

[23] Peng Gao, Jiaming Han, Renrui Zhang, Ziyi Lin, Shijie Geng, Aojun Zhou, Wei Zhang, Pan Lu, Conghui He, Xiangyu Yue, et al. LLaMA-Adapter V2: Parameter-Efficient Visual Instruction Model. *arXiv preprint arXiv:2304.15010*, 2023.

[24] Ali Goli and Amandeep Singh. Frontiers: Can Large Language Models Capture Human Preferences? *Marketing Science*, 2024.

[25] Xianda Guo, Ruijun Zhang, Yiqun Duan, Yuhang He, Chenming Zhang, Shuai Liu, and Long Chen. DriveMLLM: A Benchmark for Spatial Understanding with Multimodal Large Language Models in Autonomous Driving. *arXiv preprint arXiv:2411.13112*, 2024.

[26] Ziang Guo, Konstantin Gubernatorov, Selamawit Asfaw, Zakhar Yagudin, and Dzmitry Tsetserukou. VDT-Auto: End-to-End Autonomous Driving with VLM-Guided Diffusion Transformers. *arXiv preprint arXiv:2502.20108*, 2025.

[27] Wenyi Hong, Weihan Wang, Qingsong Lv, Jiazheng Xu, Wenmeng Yu, Junhui Ji, Yan Wang, Zihan Wang, Yuxiao Dong, Ming Ding, et al. CogAgent: A Visual Language Model for GUI Agents. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14281–14290, 2024.

[28] Yihan Hu, Jiazhi Yang, Li Chen, Keyu Li, Chonghao Sima, Xizhou Zhu, Siqi Chai, Senyao Du, Tianwei Lin, Wenhai Wang, et al. Planning-Oriented Autonomous Driving. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17853–17862, 2023.

[29] Junjie Huang, Guan Huang, Zheng Zhu, Yun Ye, and Dalong Du. BEVDet: High-Performance Multi-Camera 3D Object Detection in Bird-Eye-View. *arXiv preprint arXiv:2112.11790*, 2021.

[30] Zhijian Huang, Tao Tang, Shaoxiang Chen, Sihao Lin, Zequn Jie, Lin Ma, Guangrun Wang, and Xiaodan Liang. Making Large Language Models Better Planners with Reasoning-Decision Alignment. In *European Conference on Computer Vision*, pages 73–90. Springer, 2024.

[31] Jyh-Jing Hwang, Runsheng Xu, Hubert Lin, Wei-Chih Hung, Jingwei Ji, Kristy Choi, Di Huang, Tong He, Paul Covington, Benjamin Sapp, Yin Zhou, James Guo, Dragomir Anguelov, and Mingxing Tan. EMMA: End-to-End Multimodal Model for Autonomous Driving. *Transactions on Machine Learning Research*, 2025.

[32] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of Hallucination in Natural Language Generation. *ACM Computing Surveys*, 55(12):1–38, 2023.

[33] Bo Jiang, Shaoyu Chen, Bencheng Liao, Xingyu Zhang, Wei Yin, Qian Zhang, Chang Huang, Wenyu Liu, and Xinggang Wang. Senna: Bridging Large Vision-Language Models and End-to-End Autonomous Driving. *arXiv preprint arXiv:2410.22313*, 2024.

[34] Jinkyu Kim, Anna Rohrbach, Trevor Darrell, John Canny, and Zeynep Akata. Textual Explanations for Self-Driving Vehicles. In *European Conference on Computer Vision*, pages 563–578. Springer, 2018.

[35] Lingdong Kong, Youquan Liu, Xin Li, Runnan Chen, Wenwei Zhang, Jiawei Ren, Liang Pan, Kai Chen, and Ziwei Liu. Robo3D: Towards Robust and Reliable 3D Perception Against Corruptions. In *IEEE/CVF International Conference on Computer Vision*, pages 19994–20006, 2023.

[36] Lingdong Kong, Shaoyuan Xie, Hanjiang Hu, Lai Xing Ng, Benoit Cottereau, and Wei Tsang Ooi. RoboDepth: Robust Out-of-Distribution Depth Estimation Under Corruptions. *Advances in Neural Information Processing Systems*, 36:21298–21342, 2023.

[37] Lingdong Kong, Shaoyuan Xie, Hanjiang Hu, Yaru Niu, Wei Tsang Ooi, Benoit R. Cottereau, Lai Xing Ng, Yuexin Ma, Wenwei Zhang, Liang Pan, Kai Chen, Ziwei Liu, Weichao Qiu, Wei Zhang, Xu Cao, Hao Lu, Ying-Cong Chen, Caixin Kang, Xinning Zhou, Chengyang Ying, Wentao Shang, Xingxing Wei, Yinpeng Dong, Bo Yang, Shengyin Jiang, Zeliang Ma, Dengyi Ji, Haiwen Li, Xingliang Huang, Yu Tian, Genghua Kou, Fan Jia, Yingfei Liu, Tiancai Wang, Ying Li, Xiaoshuai Hao, Yifan Yang, Hui Zhang, Mengchuan Wei, Yi Zhou, Haimei Zhao, Jing Zhang, Jinke Li, Xiao He, Xiaoqiang Cheng, Bingyang Zhang, Lirong Zhao, Dianlei Ding, Fangsheng Liu, Yixiang Yan, Hongming Wang, Nanfei Ye, Lun Luo, Yubo Tian, Yiwei Zuo, Zhe Cao, Yi Ren, Yunfan Li, Wenjie Liu, Xun Wu, Yifan Mao, Ming Li, Jian Liu, Jiayang Liu, Zihan Qin, Cunxi Chu, Jialei Xu, Wenbo Zhao, Junjun Jiang, Xianming Liu, Ziyan Wang, Chiwei Li, Shilong Li, Chendong Yuan, Songyue Yang, Wentao Liu, Peng Chen, Bin Zhou, Yubo Wang, Chi Zhang, Jianhang Sun, Hai Chen, Xiao Yang, Lizhong Wang, Dongyi Fu, Yongchun Lin, Huitong Yang, Haoang Li, Yadan Luo, Xianjing Cheng, and Yong Xu. The RoboDrive Challenge: Drive Anytime Anywhere in Any Condition. *arXiv preprint arXiv:2405.08816*, 2024.

[38] Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient Memory Management for

Large Language Model Serving with PagedAttention. In *ACM SIGOPS Symposium on Operating Systems Principles*, pages 611–626, 2023.

[39] Boyun Li, Xiao Liu, Peng Hu, Zhongqin Wu, Jiancheng Lv, and Xi Peng. All-In-One Image Restoration for Unknown Corruption. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17452–17462, 2022.

[40] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating Object Hallucination in Large Vision-Language Models. In *Conference on Empirical Methods in Natural Language Processing*, pages 292–305, 2023.

[41] Zhiqi Li, Wenhai Wang, Hongyang Li, Enze Xie, Chonghao Sima, Tong Lu, Yu Qiao, and Jifeng Dai. BEVFormer: Learning Bird's-Eye-View Representation from Multi-Camera Images via Spatiotemporal Transformers. In *European Conference on Computer Vision*, pages 1–18. Springer, 2022.

[42] Zhiqi Li, Zhiding Yu, Shiyi Lan, Jiahan Li, Jan Kautz, Tong Lu, and Jose M Alvarez. Is Ego Status All You Need for Open-Loop End-to-End Autonomous Driving? In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14864–14873, 2024.

[43] Guibiao Liao, Jiankun Li, and Xiaoqing Ye. VLM2Scene: Self-Supervised Image-Text-LiDAR Learning with Foundation Models for Autonomous Driving Scene Understanding. In *AAAI Conference on Artificial Intelligence*, pages 3351–3359, 2024.

[44] Chin-Yew Lin. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*, pages 74–81, 2004.

[45] Benlin Liu, Yuhao Dong, Yiqin Wang, Yongming Rao, Yansong Tang, Wei-Chiu Ma, and Ranjay Krishna. Coarse Correspondences Elicit 3D Spacetime Understanding in Multimodal Language Model. *arXiv preprint arXiv:2408.00754*, 2024.

[46] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual Instruction Tuning. In *Advances in Neural Information Processing Systems*, pages 34892–34916, 2023.

[47] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved Baselines with Visual Instruction Tuning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306, 2024.

[48] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. LLaVA-NeXT: Improved Reasoning, OCR, and World Knowledge, 2024.

[49] Yinhong Liu, Han Zhou, Zhijiang Guo, Ehsan Shareghi, Ivan Vulić, Anna Korhonen, and Nigel Collier. Aligning with Human Judgement: The Role of Pairwise Preference in Large Language Model Evaluators. In *Conference on Language Modeling*, 2024.

[50] Zuyan Liu, Yuhao Dong, Yongming Rao, Jie Zhou, and Jiwen Lu. Chain-of-Spot: Interactive Reasoning Improves Large Vision-Language Models. *arXiv preprint arXiv:2403.12966*, 2024.

[51] Zuyan Liu, Yuhao Dong, Ziwei Liu, Winston Hu, Jiwen Lu, and Yongming Rao. Oryx MLLM: On-Demand Spatial-Temporal Understanding at Arbitrary Resolution. In *International Conference on Learning Representations*, 2025.

[52] Yingzi Ma, Yulong Cao, Jiachen Sun, Marco Pavone, and Chaowei Xiao. Dolphins: Multimodal Language Model for Driving. In *European Conference on Computer Vision*, pages 403–420. Springer, 2024.

[53] Yunsheng Ma, Burhaneddin Yaman, Xin Ye, Jingru Luo, Feng Tao, Abhirup Mallik, Ziran Wang, and Liu Ren. MTA: Multimodal Task Alignment for BEV Perception and Captioning. *arXiv preprint arXiv:2411.10639*, 2024.

[54] Srikanth Malla, Chiho Choi, Isht Dwivedi, Joon Hee Choi, and Jiachen Li. DRAMA: Joint Risk Localization and Captioning in Driving. In *IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1043–1052, 2023.

[55] Jiageng Mao, Yuxi Qian, Junjie Ye, Hang Zhao, and Yue Wang. GPT-Driver: Learning to Drive with GPT. *arXiv preprint arXiv:2310.01415*, 2023.

[56] OpenDriveLab. Foundation Models for Autonomous Systems, 2024. Accessed: 2024-11-11.

[57] Chenbin Pan, Burhaneddin Yaman, Tommaso Nesti, Abhirup Mallik, Alessandro G Allievi, Senem Velipasalar, and Liu Ren. VLP: Vision Language Planning for Autonomous Driving. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14760–14769, 2024.

[58] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: A Method for Automatic Evaluation of Machine Translation. In *Annual Meeting of the Association for Computational Linguistics*, pages 311–318, 2002.

[59] Tianwen Qian, Jingjing Chen, Linhai Zhuo, Yang Jiao, and Yu-Gang Jiang. nuScenes-QA: A Multi-Modal Visual Question Answering Benchmark for Autonomous Driving Scenario. In *AAAI Conference on Artificial Intelligence*, pages 4542–4550, 2024.

[60] Javier Rando, Jie Zhang, Nicholas Carlini, and Florian Tramèr. Adversarial ML Problems Are Getting Harder to Solve and to Evaluate. *arXiv preprint arXiv:2502.02260*, 2025.

[61] Enna Sachdeva, Nakul Agarwal, Suhas Chundi, Sean Roelofs, Jiachen Li, Mykel Kochenderfer, Chiho Choi, and Behzad Dariush. Rank2Tell: A Multimodal Driving Dataset for Joint Importance Ranking and Reasoning. In *IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 7513–7522, 2024.

[62] Hao Shao, Yuxuan Hu, Letian Wang, Guanglu Song, Steven L Waslander, Yu Liu, and Hongsheng Li. LMDrive: Closed-Loop End-to-End Driving with Large Language Models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15120–15130, 2024.

[63] Chonghao Sima, Katrin Renz, Kashyap Chitta, Li Chen, Hanxue Zhang, Chengen Xie, Jens Beißwenger, Ping Luo, Andreas Geiger, and Hongyang Li. DriveLM: Driving with Graph Visual Question Answering. In *European Conference on Computer Vision*, pages 256–274. Springer, 2024.

[64] Austin Stone, Ted Xiao, Yao Lu, Keerthana Gopalakrishnan, Kuang-Huei Lee, Quan Vuong, Paul Wohlhart, Sean Kirmani, Brianna Zitkovich, Fei Xia, et al. Open-World Object Manipulation Using Pre-Trained Vision-Language Models. *arXiv preprint arXiv:2303.00905*, 2023.

[65] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, et al. Scalability in Perception for Autonomous Driving: Waymo Open Dataset. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2446–2454, 2020.

[66] Xiaoyu Tian, Junru Gu, Bailin Li, Yicheng Liu, Yang Wang, Zhiyong Zhao, Kun Zhan, Peng Jia, XianPeng Lang, and Hang Zhao. DriveVLM: The Convergence of Autonomous Driving and Large Vision-Language Models. In *Conference on Robot Learning*, pages 4698–4726. PMLR, 2024.

[67] Ngoc Tran, Hieu Tran, Son Nguyen, Hoan Nguyen, and Tien Nguyen. Does BLEU Score Work for Code Migration? In *IEEE/ACM International Conference on Program Comprehension*, pages 165–176, 2019.

[68] Haoqin Tu, Chenhang Cui, Zijun Wang, Yiyang Zhou, Bingchen Zhao, Junlin Han, Wangchunshu Zhou, Huaxiu Yao, and Cihang Xie. How Many Unicorns Are in This Image? A Safety Evaluation Benchmark for Vision LLMs. *arXiv preprint arXiv:2311.16101*, 2023.

[69] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. CIDEr: Consensus-Based Image Description Evaluation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4566–4575, 2015.

[70] Junyang Wang, Yiyang Zhou, Guohai Xu, Pengcheng Shi, Chenlin Zhao, Haiyang Xu, Qinghao Ye, Ming Yan, Ji Zhang, Jihua Zhu, et al. Evaluation and Analysis of Hallucination in Large Vision-Language Models. *arXiv preprint arXiv:2308.15126*, 2023.

[71] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-VL: Enhancing Vision-Language Model's Perception of the World at Any Resolution. *arXiv preprint arXiv:2409.12191*, 2024.

[72] Shihao Wang, Zhiding Yu, Xiaohui Jiang, Shiyi Lan, Min Shi, Nadine Chang, Jan Kautz, Ying Li, and Jose M. Alvarez. OmniDrive: A Holistic Vision-Language Dataset for Autonomous Driving with Counterfactual Reasoning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22442–22452, 2025.

[73] Yue Wang, Vitor Campagnolo Guizilini, Tianyuan Zhang, Yilun Wang, Hang Zhao, and Justin Solomon. DETR3D: 3D Object Detection from Multi-View Images via 3D-to-2D Queries. In *Conference on Robot Learning*, pages 180–191. PMLR, 2022.

[74] Licheng Wen, Daocheng Fu, Xin Li, Xinyu Cai, Tao Ma, Pinlong Cai, Min Dou, Botian Shi, Liang He, and Yu Qiao. DiLu: A Knowledge-Driven Approach to Autonomous Driving with Large Language Models. In *International Conference on Learning Representations*, 2024.

[75] Dongming Wu, Wencheng Han, Tiancai Wang, Yingfei Liu, Xiangyu Zhang, and Jianbing Shen. Language Prompt for Autonomous Driving. *arXiv preprint arXiv:2309.04379*, 2023.

[76] Shaoyuan Xie, Lingdong Kong, Wenwei Zhang, Jiawei Ren, Liang Pan, Kai Chen, and Ziwei Liu. Benchmarking and Improving Bird's Eye View Perception Robustness in Autonomous Driving. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(5):3878–3894, 2025.

[77] Yichen Xie, Runsheng Xu, Tong He, Jyh-Jing Hwang, Katie Luo, Jingwei Ji, Hubert Lin, Letian Chen, Yiren Lu, Zhaoqi Leng, et al. S4-Driver: Scalable Self-Supervised Driving Multimodal Large Language Model with Spatio-Temporal Visual Representation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1622–1632, 2025.

[78] Yiran Xu, Xiaoyin Yang, Lihang Gong, Hsuan-Chu Lin, Tz-Ying Wu, Yunsheng Li, and Nuno Vasconcelos. Explainable Object-Induced Action Decision for Autonomous Vehicles. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9523–9532, 2020.

[79] Yi Xu, Yuxin Hu, Zaiwei Zhang, Gregory P Meyer, Siva Karthik Mustikovela, Siddhartha Srinivasa, Eric M Wolff, and Xin Huang. VLM-AD: End-to-End Autonomous Driving Through Vision-Language Model Supervision. *arXiv preprint arXiv:2412.14446*, 2024.

[80] 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*, 9(10):8186–8193, 2024.

[81] Jingkang Yang, Yuhao Dong, Shuai Liu, Bo Li, Ziyue Wang, Haoran Tan, Chencheng Jiang, Jiamu Kang, Yuanhan Zhang, Kaiyang Zhou, et al. Octopus: Embodied Vision-Language Programmer from Environmental Feedback. In *European Conference on Computer Vision*, pages 20–38. Springer, 2024.

[82] Zhenjie Yang, Xiaosong Jia, Hongyang Li, and Junchi Yan. A Survey of Large Language Models for Autonomous Driving. *arXiv preprint arXiv:2311.01043*, 2023.

[83] Jiayi Ye, Yanbo Wang, Yue Huang, Dongping Chen, Qihui Zhang, Nuno Moniz, Tian Gao, Werner Geyer, Chao Huang, Pin-Yu Chen, Nitesh V Chawla, and Xiangliang Zhang. Justice or Prejudice? Quantifying Biases in LLM-as-a-Judge. In *International Conference on Learning Representations*, 2025.

[84] Junwei You, Haotian Shi, Zhuoyu Jiang, Zilin Huang, Rui Gan, Keshu Wu, Xi Cheng, Xiaopeng Li, and Bin Ran. V2X-VLM: End-to-End V2X Cooperative Autonomous Driving Through Large Vision-Language Models. *arXiv preprint arXiv:2408.09251*, 2024.

[85] Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. BDD100K: A Diverse Driving Dataset for Heterogeneous Multitask Learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2636–2645, 2020.

[86] Xingcheng Zhou, Xuyuan Han, Feng Yang, Yunpu Ma, and Alois C Knoll. OpenDriveVLA: Towards End-to-End Autonomous Driving with Large Vision Language Action Model. *arXiv preprint arXiv:2503.23463*, 2025.

[87] Zewei Zhou, Tianhui Cai, Seth Z Zhao, Yun Zhang, Zhiyu Huang, Bolei Zhou, and Jiaqi Ma. AutoVLA: A Vision-Language-Action Model for End-to-End Autonomous Driv-

ing with Adaptive Reasoning and Reinforcement Fine-Tuning. *arXiv preprint arXiv:2506.13757*, 2025.

[88] Brianna Zitkovich, Tianhe Yu, Sichun Xu, Peng Xu, Ted Xiao, Fei Xia, Jialin Wu, Paul Wohlhart, Stefan Welker, Ayzaan Wahid, et al. RT-2: Vision-Language-Action Models Transfer Web Knowledge to Robotic Control. In *Conference on Robot Learning*, pages 2165–2183. PMLR, 2023.