IntelliCockpitBench: A Comprehensive Benchmark to Evaluate VLMs for Intelligent Cockpit

Liang Lin[1*], **Siyuan Chai**[1*], **Jiahao Wu**[1*], **Hongbing Hu**[1*†], **Xiaotao Gu**[1]
Hao Hu[2], **Fan Zhang**[1], **Wei Wang**[3,4†], **Dan Zhang**[4†]
[1]Zhipu AI; [2]Mercedes-Benz; [3]Nankai University; [4]Tsinghua University
https://github.com/Lane315/IntelliCockpitBench/

Abstract

The integration of sophisticated Vision-Language Models (VLMs) in vehicular systems is revolutionizing vehicle interaction and safety, performing tasks such as Visual Question Answering (VQA). However, a critical gap persists due to the lack of a comprehensive benchmark for multimodal VQA models in vehicular scenarios. To address this, we propose IntelliCockpitBench, a benchmark that encompasses diverse automotive scenarios. It includes images from front, side, and rear cameras, various road types, weather conditions, and interior views, integrating data from both moving and stationary states. Notably, all images and queries in the benchmark are verified for high levels of authenticity, ensuring the data accurately reflects real-world conditions. A sophisticated scoring methodology combining human and model-generated assessments enhances reliability and consistency. Our contributions include a diverse and authentic dataset for automotive VQA and a robust evaluation metric aligning human and machine assessments. IntelliCockpitBench is open-sourced and publicly available at https://github.com/Lane315/IntelliCockpitBench.

1 Introduction

In recent years, with the advancement of Visual Language Models (VLMs) (Liu et al., 2023; Bai et al., 2023; Wang et al., 2023a; Zhang et al., 2025a), intelligent cockpit technology has made significant progress, becoming an important interface for the next generation of human-computer interaction. Subsequently, benchmarks like DriveBench (Xie et al., 2025) and NuScenes-QA (Qian et al., 2024) have been proposed to evaluate the visual question-answering (VQA) capabilities in autonomous driving scenarios. Even so, these benchmarks remain primarily focused on decision-making scenarios such as autonomous driving and do not adequately consider non-decision-making scenarios aimed at enhancing user experience and interaction. This has significant limitations in the research field. **Limitation 1:** the lack of comprehensive benchmarks specifically designed to evaluate the performance of VLMs in non-decision-making scenarios within intelligent cockpits. **Limitation 2:** existing GPT-based (Hurst et al., 2024; Zhang et al., 2025b) automatic evaluation methods typically rely on uniform assessment standards, which overlook the specific nature and requirements of different queries. This further emphasizes the necessity of developing evaluation benchmarks tailored to different queries types.

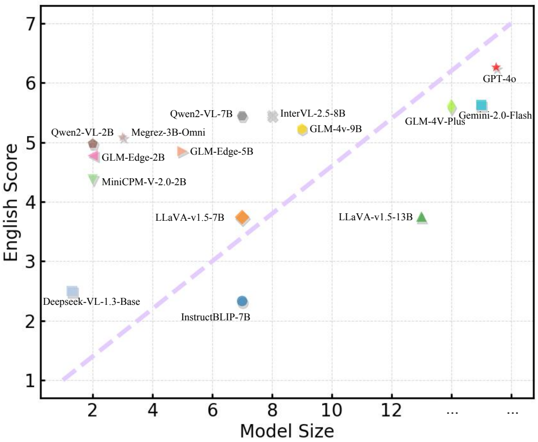


Figure 1: The relationship between model size and score in English queries across various VLMs on IntelliCockpitBench. Notable models such as GPT-4o (Hurst et al., 2024) and Gemini-2.0-Flash (Team et al., 2023) are distinguished by their superior performance despite larger sizes. The dotted line represents an estimated trend indicating the positive correlation between model size and performance.

To address these limitations, we propose a comprehensive benchmark named IntelliCockpitBench to evaluate VLMs for intelligent cockpits. This benchmark includes

[*]Equal contribution.
[†]Corresponding authors.

Findings of the Association for Computational Linguistics: ACL 2025, pages 15453–15475
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics

a diverse collection of images captured from front, side, and rear cameras, encompassing various road types and weather conditions to provide a comprehensive external perspective. Additionally, IntelliCockpitBench features interior images to reflect the complexity of the in-vehicle environment. The curated dataset also integrates data from both moving and stopping vehicle states, ensuring a thorough representation of real-world scenarios (Wang et al., 2024b). Taking into account the scenarios of visual information augmented, we have also implemented data augmentation techniques to ensure the robustness of IntelliCockpitBench in various unexpected situations. All queries in our dataset are collected through driver surveys and generalized using GPT-4o (Hurst et al., 2024) to ensure their authenticity and diversity. Note that all included images and queries are verified for high levels of authenticity and have undergone human review, which ensures that the data accurately reflects real-world driving scenarios.

Furthermore, we design three key LLM-as-a-judge methods including Chain-of-Thought Reasoning, Multi-dimensional Variance Analysis, and Rule-Calibrated Referencing. This evaluation method not only defines different evaluation metrics for various queries but also assigns importance scores to these metrics. Additionally, it utilizes Chain-of-Thought to generate explanations and final ratings, ensuring both high reliability and interpretability. As shown in Figure 1 and Table 3, we evaluate 15 VLMs and our experiments reveal that current VLMs perform poorly when confronted with augmented visual images and queries requiring deep reasoning. Therefore, it is essential to enhance VLMs' capabilities in accurate visual localization and multi-step reasoning queries.

Overall, our key contributions are as follows:

- We create a comprehensive benchmark, IntelliCockpitBench, to evaluate the capabilities of VLMs for the intelligent cockpit, featuring 5 intelliCockpit query types, 38 driving scenarios, 10+ question formats, $16,154$ queries, over $7,622$ images, and 20 evaluation metrics.
- We propose 3 innovative LLM-as-a-judge evaluation methods including Chain-of-Thought Reasoning, Multi-dimensional Variance Analysis, and Rule-Calibrated Referencing to enhance the reliability and interpretability of evaluation.
- We evaluate 15 open-source and closed-source VLMs and find that all models perform poorly on the IntelliCockpitBench, especially with augmented visuals and complex reasoning queries, highlighting the need for improved visual localization and reasoning in VLMs.

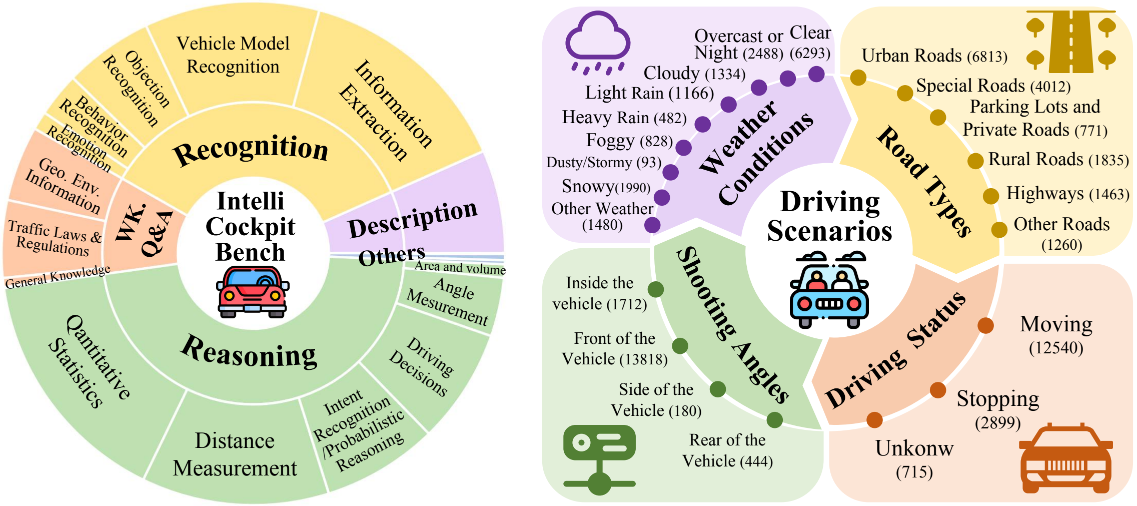
2 Related Work

2.1 Vision-Language Models

The success of Large Language Models (LLMs) (Touvron et al., 2023; Team et al., 2023; GLM et al., 2024; Zhang et al., 2025a) has significantly advanced VLMs. BLIP (Liu et al., 2024a) employs GPT-4 to generate instruction-following data for vision-language tuning, and its learning paradigm and instruction-tuning corpus have been widely adopted in subsequent research (Chen et al., 2025, 2024a). Over the past year, numerous open-source VLMs have gained recognition, including the LLaVA series (Liu et al., 2024a,c,b), MiniGPT-4 (Zhu et al., 2023), Vision-LLM (Wang et al., 2024c), Qwen-VL (Bai et al., 2023; Wang et al., 2024a), CogVLM (Wang et al., 2023a), Intern-VL (Chen et al., 2024b; Dong et al., 2024), and others (Chen et al., 2023; Peng et al., 2023; Wang et al., 2023b). Although these models are generally aimed at standard VQA and various broad applications, there is still a clear gap in their use within smart cockpit settings. Regarding this, we propose IntelliCockpitBench encompassing **5 query types** and **4 scenarios** in Figure 2.

2.2 Multimodal Datasets

Existing vision-and-language benchmarks for intelligent vehicles primarily focus on two distinct usage scenarios: autonomous driving and intelligent cockpit interactions. Representative autonomous driving benchmarks include DriveLM (Sima et al., 2023), NuScenes-QA (Qian et al., 2024), and DriveBench (Xie et al., 2025), all of which primarily address VQA tasks centered on perception and control aspects relevant to driving. These benchmarks generally employ a query-based taxonomy and are evaluated using large language models, often relying solely on GPT-based assessment methods. In contrast, SuperCLUE-o (Xu et al., 2020) and our proposed IntelliCockpitBench are designed for intelligent cockpit environments, where the primary focus is on enhancing the passenger experience rather than influencing vehicle operation decisions.



(a) Distribution of query types.

(b) Distribution of driving scenarios.

Figure 2: A comprehensive taxonomy of query types and scenarios in VLMs within IntelliCockpitBench. "WK." denotes World Knowledge. "Geo. Env." denotes Geospatial environmental.

Table 1 provides a detailed comparison of these benchmarks. Notably, IntelliCockpitBench introduces several novel aspects. First, it adopts a taxonomy structure that integrates both queries and driving scenarios, accounting for diverse real-world conditions such as varying road types and weather. Second, unlike prior works that predominantly rely on GPT-based evaluations, IntelliCockpitBench incorporates a multi-faceted evaluation methodology, combining rule-based assessment, Chain-of-Thought (CoT) reasoning, multi-dimensional variance analysis, and rule-calibrated referencing. This approach enables a more comprehensive evaluation of model performance across both edge and large-scale language models. Additionally, unlike earlier benchmarks with imbalanced data distributions, IntelliCockpitBench ensures balanced representation across categories, facilitating fairer evaluation. These advancements position IntelliCockpitBench as a significantly more effective tool for benchmarking models designed for intelligent cockpit applications.

3 IntelliCockpitBench

In this section, we introduce an overview of the data composition, the dataset construction, and the evaluation paradigm of IntelliCockpitBench.

3.1 Dataset Composition

To ensure the authenticity and diversity of the curated dataset, we first collect images and queries that are sourced from real-world driving scenarios. We then propose a comprehensive taxonomy for VLMs' driving queries based on real-driver queries to conduct a systematic evaluation. As illustrated in Figure 2, from simple descriptions to complex reasoning, these queries are divided into **5 dimensions**: description, recognition, world knowledge Q&A, reasoning, and others. The detailed explanation of each query is provided in Appendix A.1.

In addition, to thoroughly evaluate the adaptability and robustness of VLMs given the complexity and variability of real-world driving scenarios, we categorize and summarize these scenarios into **4 categories** including weather conditions, road types, driving status, and shooting angles), **38 meta-categories**, and a total of **7,622 images**. We provide a detailed taxonomy and definition for these four driving scenarios in Appendix A.2.

3.2 Dataset Construction

This subsection delineates the construction process of the dataset, encompassing three primary stages: image generation, query generation, and answer generation, as illustrated in Figure 3.

3.2.1 Image Generation

Overall, our image data sources can be classified into two principal categories. The first category encompasses partial data collection from publicly available datasets, including nuScene, the YawDD (Abtahi et al., 2014), and the Drive&Act (Martin et al., 2019) dataset. The second category, representing the primary source of

Table 1: Comparison between IntelliCockpitBench and related benchmarks. "AutoDri." denotes Autonomous Driving, "IntCock." denotes Intelligent Cockpit, "DataDis." denotes Data Distribution, "EvaMTpyes." denotes Evaluation Model Types, "EvaModels." denotes Evaluation Models.

Benchmark	Usage Scenarios	Taxonomy	#VQA Pairs	DataDis.	GPT Evaluation	EvaMTpyes.	#EvaModels	Available
DriveLM	AutoDri.	Queries	15,480	Imbalance	GPT	Above 3B	5	✗
NuScenes-QA	AutoDri.	Queries	83,337	Imbalance	GPT	Above 3B	2	✓
DriveBench	AutoDri.	Queries	20,498	Balance	GPT	Above 3B	12	✓
SuperCLUE-O	IntCock.	-	-	-	GPT + Rules + CoT + Multi-Dimensions	Edge (1B–3B) + (Above 3B)	-	✗
IntelliCockpitBench	IntCock.	Queries + Driving Scenarios	16,154	Balance	GPT + Rules + CoT + Multi-Dimensions	Edge (1B–3B) + (Above 3B)	15	✓

our dataset, comprises over 100 meticulously selected driving videos obtained from video-sharing platforms. Download data for academic research only. These videos are rigorously chosen based on a carefully defined taxonomy of driving scenarios (refer to Appendix A.2). Subsequently, we systematically sample frames from the collected videos at consistent intervals of every 15 second, culminating in an extensive dataset consisting of $7,622$ images. All images have undergone a de-identification process to mask faces and license plate numbers. Considering the substantial impact that image quality has on the performance of VLMs, our dataset intentionally includes images of various resolutions.

In addition to designing and screening images under normal driving conditions, we consider scenarios where visual information degrades, such as weather-induced image quality degradation (rain or fog), changes in lighting (overexposure), and camera malfunctions (image distortion, obstruction, or misalignment). A total of **190 images** are collected to validate the robustness and reliability of VLMs under various unforeseen circumstances.

3.2.2 Query Generation

Most existing VQA benchmarks are limited in the diversity of questioning types (Xie et al., 2025; Xu et al., 2017), failing to fully represent the wide spectrum of human conversations. In contrast, the questioning set in IntelliCockpitBench has been carefully curated to include a broad range of categories. Figure 8 illustrates the distributions of the questioning type. Questioning types include 'what', 'who', 'how', 'when', and 'where'. We also expand scopes of type to include interrogatives like 'why', 'which', 'is/are', and 'does/do'. This expansion enhances the diversity and better reflects the natural style of human dialogues.

Real-driver Query Generation. Due to the lack of authenticity in queries generated directly based on images using GPT-4o (Hurst et al., 2024), we obtain real intelligent cockpit queries by recruiting **100 drivers**. Each driver provides **100 queries** they might encounter in driving scenarios related to visual information, resulting in a total of **10,000 real-driver queries**. To ensure the diversity of queries, we use GPT-4o to generalize them. Specifically, we first leverage the classification results of the questioning type and then perform random sampling from the collected real dataset as a few-shot input to generate new queries. The detailed query generation prompt is in Appendix A.9.

Human Check. The content generated by the GPT-4o (Hurst et al., 2024) is then subjected to manual inspection to ensure that both the image and the query are of high quality and accurately represent realistic scenarios, which are conducted in two stages. Initially, we ask annotators to evaluate whether the generated queries meet the five specific criteria listed in Table 4. Queries that do not meet these criteria will be manually modified, and if modification is not feasible, they will be discarded. The establishment and implementation of refusal strategies for VLMs are crucial, as they can effectively prevent misinformation, protect user privacy, and ensure that the generated content adheres to ethical and legal standards. Subsequently, for the queries that pass the initial evaluation, annotators further determine whether the query needs to be refused an answer, as outlined in the rejection strategy presented in Appendix A.3. We provide details of human checks in Appendix A.4.

3.2.3 Answer Generation

Following the generation of high-quality images and realistic queries, the next step involves con-

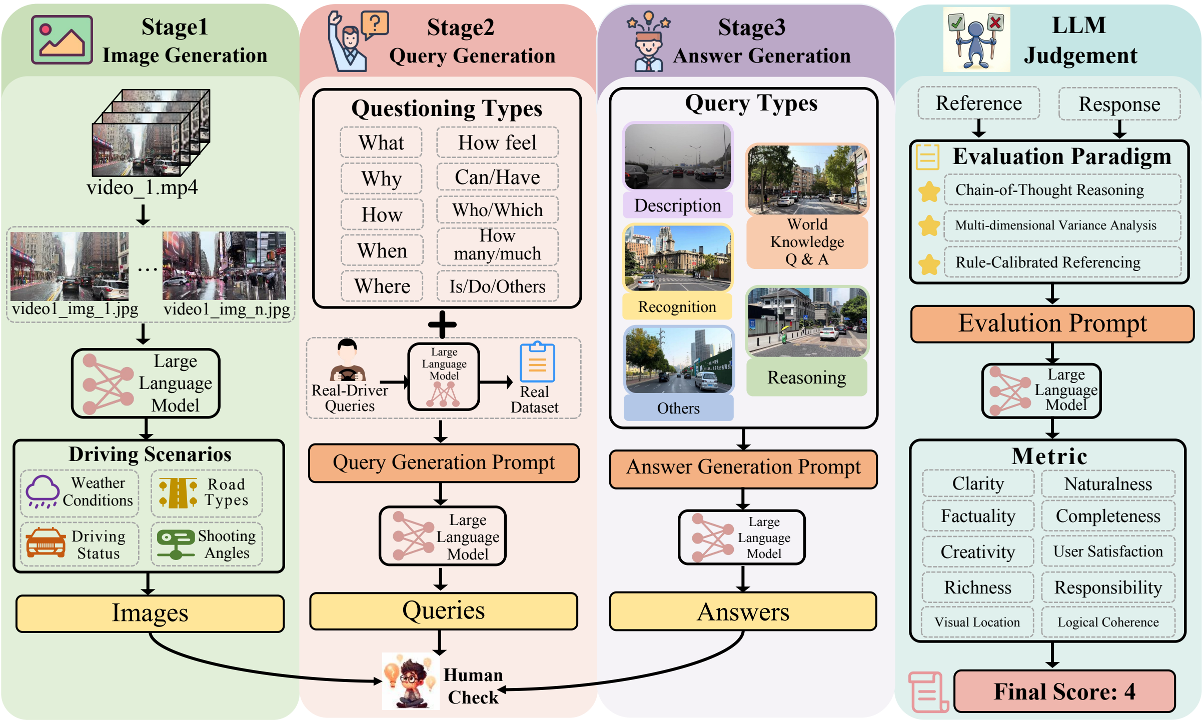


Figure 3: Architecture of the proposed IntelliCockpitBench. Dataset construction involves three steps: **1) Image Generation**, creating driving scenario images using video generation techniques; **2) Query Generation**, generating multiple types of intellicockpit queries using LLMs and real-driver queries; **3) Answer Generation**, producing corresponding answers based on different intellicockpit queries; The last module is LLM Judgement, scoring multiple dimensions of the answers using evaluation paradigms based on chain-of-thought reasoning, multi-dimensional variance analysis, and rule-based calibration, ultimately providing a comprehensive score.

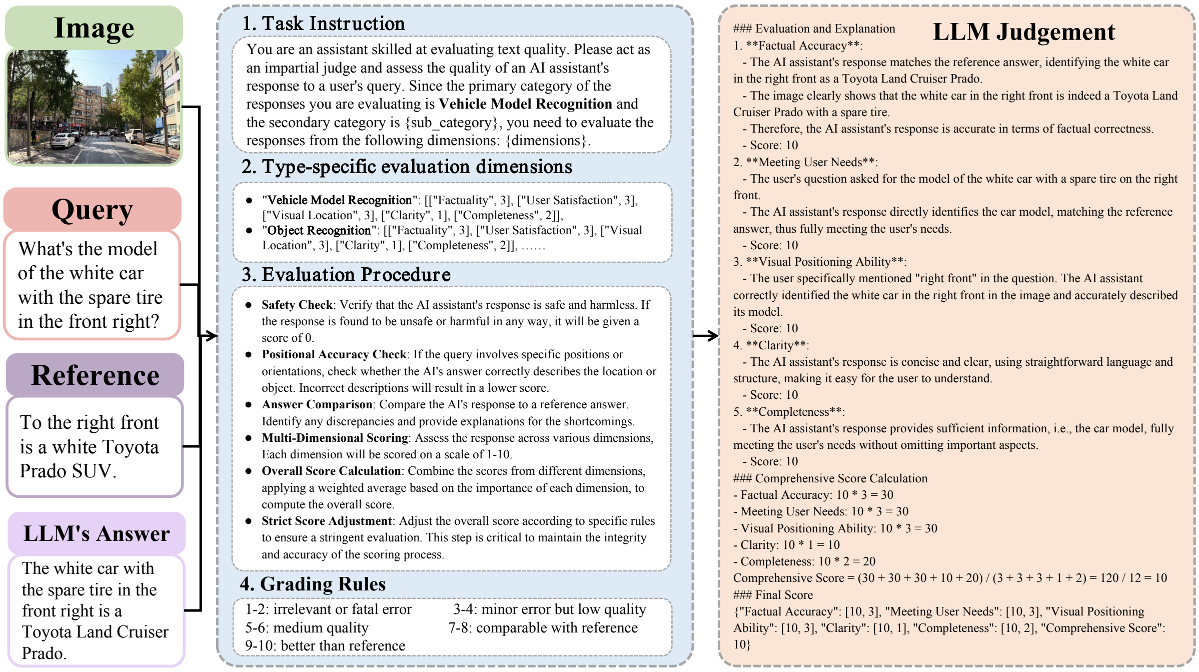
structing accurate answers. Specifically, the previously generated images and queries, along with the VLMs' queries categorization system, are input into GPT-4o. This process enables the model to produce a clear answer, a concise rationale, and the corresponding query labels. We provide an answer generation prompt in Appendix A.9. The final outputs are then manually verified to ensure their authenticity and accuracy. First, we instruct annotators to confirm that the answer correctly addresses the query based on the image. Next, they ensure that the classification of both the query and the image aligns with the established categorization system. If any inaccuracies are identified, the annotators manually revise the answers. Note that all VQA pairs generated in IntelliCockpitBench undergo a rigorous cross-validation process (see Appendix A.4) to ensure their accuracy and adherence to the classification system.

3.3 Evaluation Paradigm

To effectively evaluate the quality of VLMs' responses, IntelliCockpitBench utilizes GPT-4o-Mini-2024-07-18 (Hurst et al., 2024) as the primary evaluator to analyze and grade responses in accordance with established practices (Zheng et al., 2023). Nonetheless, a significant design space in VQA remains unexplored, particularly regarding prompting strategies, score calibration, critique explainability, and evaluation dimensions. To address these gaps, we develop a rule-based evaluation methodology using Language Models as Judges (LLM-as-a-judge) that incorporates three principal approaches: Chain-of-Thought Reasoning, Multi-dimensional Variance Analysis, and Rule-Calibrated Referencing. Detailed prompts for rule-based evaluation are provided in Appendix A.9, and an illustrative example is shown in Figure 4.

Chain-of-Thought Reasoning. When leveraging LLM-as-a-Judge, IntelliCockpitBench employs point-wise grading to assess the quality of responses. The inputs include the image, the query associated with the image, the model's response, and a human-curated reference answer. The output consists of a multi-dimensional analytical explanation alongside a final rating on a 1 to 10 scale.

Multi-dimensional Variance Analysis. Given the diverse nature and characteristics of different queries, applying a uniform standard to all re-



What's the model of the white car with the spare tire in the front right?

To the right front is a white Toyota Prado SUV.

The white car with the spare tire in the front right is a Toyota Land Cruiser Prado.

1. Task Instruction

You are an assistant skilled at evaluating text quality. Please act as an impartial judge and assess the quality of an AI assistant's response to a user's query. Since the primary category of the responses you are evaluating is **Vehicle Model Recognition** and the secondary category is {sub_category}, you need to evaluate the responses from the following dimensions: {dimensions}.

2. Type-specific evaluation dimensions

- "**Vehicle Model Recognition**": [["Factuality", 3], ["User Satisfaction", 3], ["Visual Location", 3], ["Clarity", 1], ["Completeness", 2]],
- "**Object Recognition**": [["Factuality", 3], ["User Satisfaction", 3], ["Visual Location", 3], ["Clarity", 1], ["Completeness", 2]], ……

3. Evaluation Procedure

- **Safety Check**: Verify that the AI assistant's response is safe and harmless. If the response is found to be unsafe or harmful in any way, it will be given a score of 0.
- **Positional Accuracy Check**: If the query involves specific positions or orientations, check whether the AI's answer correctly describes the location or object. Incorrect descriptions will result in a lower score.
- **Answer Comparison**: Compare the AI's response to a reference answer. Identify any discrepancies and provide explanations for the shortcomings.
- **Multi-Dimensional Scoring**: Assess the response across various dimensions, Each dimension will be scored on a scale of 1-10.
- **Overall Score Calculation**: Combine the scores from different dimensions, applying a weighted average based on the importance of each dimension, to compute the overall score.
- **Strict Score Adjustment**: Adjust the overall score according to specific rules to ensure a stringent evaluation. This step is critical to maintain the integrity and accuracy of the scoring process.

4. Grading Rules

1-2: irrelevant or fatal error
3-4: minor error but low quality
5-6: medium quality
7-8: comparable with reference
9-10: better than reference

LLM Judgement

```
### Evaluation and Explanation
1. **Factual Accuracy**:
   - The AI assistant's response matches the reference answer, identifying the white car in the right front as a Toyota Land Cruiser Prado.
   - The image clearly shows that the white car in the right front is indeed a Toyota Land Cruiser Prado with a spare tire.
   - Therefore, the AI assistant's response is accurate in terms of factual correctness.
   - Score: 10
2. **Meeting User Needs**:
   - The user's question asked for the model of the white car with a spare tire on the right front.
   - The AI assistant's response directly identifies the car model, matching the reference answer, thus fully meeting the user's needs.
   - Score: 10
3. **Visual Positioning Ability**:
   - The user specifically mentioned "right front" in the question. The AI assistant correctly identified the white car in the right front in the image and accurately described its model.
   - Score: 10
4. **Clarity**:
   - The AI assistant's response is concise and clear, using straightforward language and structure, making it easy for the user to understand.
   - Score: 10
5. **Completeness**:
   - The AI assistant's response provides sufficient information, i.e., the car model, fully meeting the user's needs without omitting important aspects.
   - Score: 10
### Comprehensive Score Calculation
- Factual Accuracy: 10 * 3 = 30
- Meeting User Needs: 10 * 3 = 30
- Visual Positioning Ability: 10 * 3 = 30
- Clarity: 10 * 1 = 10
- Completeness: 10 * 2 = 20
Comprehensive Score = (30 + 30 + 30 + 10 + 20) / (3 + 3 + 3 + 1 + 2) = 120 / 12 = 10
### Final Score
{"Factual Accuracy": [10, 3], "Meeting User Needs": [10, 3], "Visual Positioning Ability": [10, 3], "Clarity": [10, 1], "Completeness": [10, 2], "Comprehensive Score": 10}
```

Figure 4: An exemplar scoring process of IntelliCockpitBench on vehicle model recognition category.

sponses is inappropriate. To address this, we propose a multi-dimensional scoring approach that tailors evaluation criteria to the specific query type, providing a more detailed and structured analysis. Specifically, we define distinct evaluation dimensions and importance scores tailored to each query type. For example, in the case of descriptive queries, factuality should be prioritized, with completeness considered secondary. Consequently, the importance score for factuality is higher than that for completeness. We provide detailed definitions and settings of dimensions in Appendix A.9.

Rule-Calibrated Referencing. We provide a high-quality reference answer, which is primarily generated by GPT-4o and modified by human annotators to ensure its correctness and improve its quality. To guide the evaluator in comparing the answer with the reference and generating more controllable scores, we provide detailed grading rules that explain the relationship between score intervals and the quality of the answer compared to the reference. Additionally, we established a reference answer with a score of 8 as a benchmark for evaluation within a maximum score of 10.

4 Experiment

In this section, we conduct extensive benchmark experiments and analyses in IntelliCockpitBench, providing detailed discussions that lead to our observations and conclusions step by step.

4.1 Consistency Evaluation

To validate the alignment of the evaluation paradigm of IntelliCockpitBench with human judgment, we conduct extensive human evaluations on selected queries. Specifically, we use GPT-4o-Mini-2024-07-18 as our scoring model due to its superior accuracy and consistency in natural language processing tasks. Evaluators were instructed to analyze the model's answers and provide scores based on predefined dimensions in Appendix A.9.

To align the consistency between the scores generated by GPT-4o-Mini with those labeled by humans, we assess consistency using the following three metrics: **Sample-level Pearson Correlation:** Since each query defines different evaluation dimensions and human judges also score each dimension, we first calculate the Pearson correlation coefficient for each sample and then compute the mean as the sample-level correlation. **System-level Pearson Correlation:** This metric assesses the correlation at the system level by calculating the Pearson coefficient between the average scores at the sample-level given by human judges and model judges to the LLM. **Pairwise Agreement (w/o ties)**: For each response, scores from human judges and model judges are converted into pairwise comparisons, with ties excluded.

We also compare a modified version of the evaluation prompts used in MT-Bench (Zheng et al., 2023) as a general evaluation with our rule-based

Table 2: Comparison on human agreement between different judging methods on sampled `IntelliCockpitBench`, rated by GPT-4o. The best performance is shown in **bold**.

Metric	Method	Overall	Description	Recognition	World Knowledge Q&A	Reasoning	Others
Sample-level Pearson	ours	0.80	0.92	0.78	0.67	0.82	0.96
System-level Pearson	general	0.64	0.53	0.59	0.63	0.71	0.50
	ours	**0.93**	**0.93**	**0.90**	**0.95**	**0.94**	**0.92**
Pairwise Agreement (w/o tie)	general	0.75	0.65	0.75	0.69	0.79	0.69
	ours	**0.93**	**0.97**	**0.91**	**0.92**	**0.95**	**0.97**

Table 3: Performance evaluation of various VLMs on the `IntelliCockpitBench` for different English and Chinese VQA intelliCockpit question types. "Des." denotes Description, "Rec." denotes Recognition, "Wk-QA" denotes World Knowledge Q&A, "Rea." denotes Reasoning. Underline indicates the best results within open-source and closed-source categories, while **bold** signifies the best results among all open-source and closed-source options.

Model	**Size**	**Type**	**GPU Usage (MiB)**	**Driving Questions (EN)**						**Driving Questions (CH)**					
				Overall	Des.	Rec.	Wk-QA	Rea.	Others	**Overall**	Des.	Rec.	Wk-QA	Rea.	Others
DeepSeek-VL-base (Lu et al., 2024)	1.3B	open	5,284	3.47	3.96	3.20	4.10	3.47	3.08	2.50	2.48	2.20	3.12	2.59	3.01
GLM-Edge-V (Hong et al., 2024)	2B	open	4,566	4.34	4.80	4.51	4.96	4.02	4.30	4.78	5.74	4.73	5.51	4.52	5.15
Qwen2-VL (Wang et al., 2024a)	2B	open	28,300	4.63	4.78	4.85	5.25	4.31	4.52	4.98	6.03	5.19	5.69	4.51	5.44
MiniCPM-V-2.0 (Yao et al., 2024)	2.8B	open	9,098	4.02	3.96	4.13	4.61	3.81	4.02	4.38	5.33	4.47	5.04	4.03	3.81
Megrez-Omni (Li et al., 2025)	3B	open	10,854	4.06	3.59	4.03	4.78	4.00	3.67	5.09	5.97	5.02	5.85	4.84	5.53
GLM-Edge-V (Hong et al., 2024)	5B	open	10,152	4.51	5.19	4.63	5.18	4.17	4.43	4.85	5.95	4.62	5.66	4.68	5.49
InstructBLIP (Dai et al., 2023)	7B	open	20,456	3.83	4.08	3.46	4.44	3.94	2.96	2.33	4.05	2.17	2.72	2.13	2.04
Qwen2-VL (Wang et al., 2024a)	7B	open	39,800	5.17	5.85	5.21	6.11	4.83	5.15	5.45	6.31	5.64	6.44	4.95	5.83
LLaVA-v1.5 (Liu et al., 2024b)	7B	open	16,024	4.09	4.61	3.52	5.01	4.25	3.76	3.74	4.26	3.29	4.60	3.81	4.31
InternVL-2.5 (Chen et al., 2024b)	8B	open	24,558	5.09	5.83	5.02	5.96	4.85	4.80	5.46	6.74	5.43	6.39	5.09	5.67
GLM-4V (Hong et al., 2024)	9B	open	28,578	4.85	5.61	4.89	5.78	4.52	4.43	5.23	5.87	5.33	6.07	4.87	5.62
LLaVA-v1.5 (Liu et al., 2024b)	13B	open	28,822	4.24	4.67	3.73	5.26	4.33	3.88	3.75	4.61	3.43	4.66	3.66	4.12
GLM-4V-Plus (Hong et al., 2024)	-	closed	-	5.32	6.05	5.28	6.33	5.01	5.42	5.61	6.40	5.55	6.60	5.31	6.12
GPT-4o (Hurst et al., 2024)	-	closed	-	**5.81**	**6.36**	**5.91**	**6.77**	**5.45**	**5.70**	**6.26**	**7.37**	**6.27**	**7.26**	**5.88**	**6.27**
Gemini-2.0-Flash (Team et al., 2023)	-	closed	-	5.34	5.86	5.38	6.29	5.02	**5.70**	5.63	6.49	5.72	6.46	5.25	6.03

calibration evaluation method. The prompt for general evaluation is in appendix A.9. As presented in Table 2, results show that our pointwise multi-dimensional rules-calibrated LLM-as-a-judge method performs best, particularly on the Sample-level Pearson metric and the Pairwise Agreement (w/o tie) metric, thereby substantiating the excellent agreement with human judges. The reasons are as follows: **1)** The nature and characteristics of the driving questions in VLMS vary, making it inappropriate to apply a unified evaluation standard to all questions. **2)** Our method integrates the chain-of-thought reasoning approach to generate explanations and final scores, ensuring high reliability and interpretability. Furthermore, We plot the cumulative distribution of the human judge, general judge, and rule-calibrated judge in Figure 9 to show that the rule-calibration judge has a narrower gap to human evaluation's cumulative distribution.

4.2 `IntelliCockpitBench` Evaluation

Based on the validity of scoring and the comprehensive capabilities of `IntelliCockpitBench`, we systematically assess a diverse set of VLMs.

Result Analysis of Closed Models. As shown in Table 3 and Table 5, main results indicate that most VLMs perform poorly on `IntelliCockpitBench`, achieving an average score of only 4.58. In the analysis of our experiment, we observe that the closed-source models (GLM-4V-Plus, GPT-4o, and Gemini) consistently outperformed open-source models in both intellicockpit query performance metrics (EN and CH) and road type scenarios (EN). Specifically, GPT-4o demonstrates the highest overall performance in both English and Chinese driving questions, with exceptional performance in reasoning (Rea.), world knowledge Q&A (Wk-QA), and other driving questions categories, achieving scores of 6.77 and 7.26 respectively in these questions.

Result Analysis of Open-sourced Models.

Qwen2-VL (7B) and InternVL-2.5 (8B) are the top performers. Qwen2-VL achieves the highest scores in both overall intellicockpit query performance in Chinese (CN) with a score of 5.45 and in the special roads category for English road types, scoring 5.48. Meanwhile, InternVL-2.5 demonstrates strong performance across various English query, achieving an overall score of 5.09, including high scores in the reasoning (5.96) and urban roads categories (4.85). Notably, the larger open-source models (sizes 8B and 13B) do not consistently outperform smaller models (sizes 2B to 7B), suggesting that model architecture and training data might play more crucial roles than mere parameter size in determining query-specific performance. We follow the default open-source code to evaluate and show the model's GPU usage as a reference.

In particular, we observe that InstructBLIP, with a parameter size of 10B, performed worse on this dataset compared to smaller models (5B parameters and below). This may be due to the shorter training duration of InstructBLIP. Additionally, InstructBLIP score lower on the world knowledge question-answering queries, likely because the model is exposed to less driving scenario-related data during training.

Result Analysis of Query Types. Moreover, models of all sizes seem to outperform in Wk-QA questions compared to other categories of questions. This might be attributed to the fact that Wk-QA questions primarily evaluate the knowledge capacity of the models, and the answers to such questions are typically more singular. But for reasoning problems, especially in driving decision-making, the accuracy is notably low. This not only requires the model to have strong visual localization capabilities but also demands robust reasoning abilities. As illustrated in Figure 6, we provide the failure cases generated by advanced GPT-4o for better understanding.

4.3 Data Augmentation

Description. In real driving scenarios, the clarity of images can often not be guaranteed due to various reasons such as lighting brightness, shooting distortion, radar imaging (no color), low-pixel cameras, vehicle movement, camera occlusion, and exposure. To evaluate the robustness of VLMs in these scenarios, we employ data augmentation techniques including Clear (reduced brightness), Distorted (distortion), Grayscale (removal of image color), Low Resolution, Motion Blur, SnowEffect, and Overexposed to construct abnormal image data. We select a total of **190** images from the entire dataset, with the original images, questions, and GPT-4o's responses serving as the control group, and the augmented images, questions, and GPT-4o's responses as the experimental group. The specific augmentations and their configurations are in appendix A.6.

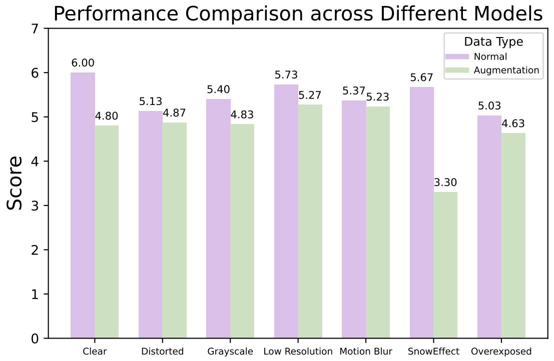


Figure 5: Comparison of scores between normal and augmented data types across various image conditions.

Result Analysis. The experimental results are shown in Figure 5, the key findings are: **1)** SnowEffect (simulating lens obstruction) have the greatest impact on the model's performance, with the score dropping from 5.67 to 3.30 (−2.37). This indicates that the model's recognition ability significantly decreases when the lens is partially obstructed. **2)** The effects of Overexposed at 4.63 (−0.4), Grayscale at 4.83 (−0.57), Clear (reduced brightness) at 4.80 (−1.2), and Low Resolution at 5.27 (−0.46) show that the model is quite sensitive to changes in lighting, color, and resolution. **3)** Under Motion Blur at 5.27 (−0.14) and Distorted (image distortion) at 4.87 (−0.26), the model still maintain good robustness, showing less impact. These results provide important references for future improvements of the model. For example, optimizing the model in terms of occlusion, lighting variations, color, and resolution to enhance the overall robustness and adaptability of the model.

4.4 Case Study

To gain a deeper understanding of VLMs' performance and robustness, we conduct case studies and choose a specific category for an in-depth case analysis focusing on reasoning questions, with a detailed examination of the scenario depicted in Figure 6 (d). **Reasoning query:** This requires the model to accurately identify the image con-

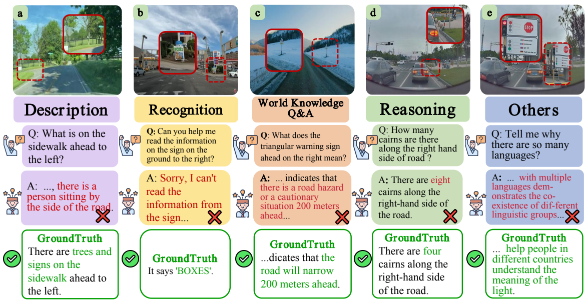

a
b
c
d
e
Description
Recognition
World Knowledge Q&A
Reasoning
Others
Q: What is on the sidewalk ahead to the left?
Q: Can you help me read the information on the sign on the ground to the right?
Q: What does the triangular warning sign ahead on the right mean?
Q: How many cairns are there along the right hand side of road ?
Q: Tell me why there are so many languages?
A: ..., there is a person sitting by the side of the road.
A: Sorry, I can't read the information from the sign...
A: ... indicates that there is a road hazard or a cautionary situation 200 meters ahead...
A: There are eight cairns along the right-hand side of the road.
A: ... with multiple languages dem-onstrates the co-existence of dif-ferent linguistic groups...
GroundTruth There are trees and signs on the sidewalk ahead to the left.
GroundTruth It says 'BOXES'.
GroundTruth ...dicates that the road will narrow 200 meters ahead.
GroundTruth There are four cairns along the right-hand side of the road.
GroundTruth ... help people in different countries understand the meaning of the light.

Figure 6: Bad cases generated by GPT-4 across five query categories. Each category presents a question and the model's generated answer is compared against the ground truth. Visual elements within each image are highlighted to indicate relevant information. Correct model responses are marked with a check, and incorrect responses are marked with a cross.

tent based on instructions and make correct conclusions based on the scenario's knowledge. **Analysis:** However, GPT-4o incorrectly identified the number of cairns on the right-hand side of the road. The model's response of “eight cairns” deviated significantly from the actual count of “four cairns”. This error indicates a need for improvement in the model's reasoning capabilities, particularly in object counting when the objects are similar in appearance and evenly spaced. **Potential improvement:** Providing more diverse and extensive training data is essential for fine-tuning VLMs, specifically targeting scenarios that require precise counting and complex visual differentiation.

4.5 Fine-Grained Analysis of Error Types

In order to understand the specific error types in more detail, we analyze from a more fine-grained perspective and dig deeper into the root causes of the errors.

As shown in Table 7 (Appendix A.7), we identify five major categories of error causes. Table 8 (Appendix A.7) presents the average distribution of these error types across different models.

Factual Errors. These involve mistakes in object recognition, quantity counting, visual positioning, and spatial distance when the model's interpretations don't align with the actual image content. For example, in the badcase where the input question was "How many cars are in our lane ahead?" the reference answer stated there was one car, but the model response claimed there were approximately five cars, highlighting a clear quantitative error and deficiency in visual positioning and environmental understanding.

Information Quality Errors. Key issues include incomplete, vague, or superficial responses, lack of depth, excessive digression, and omission of critical information. An instance is the response to "How steep is that slope?" where the model provided an overly complex and lengthy explanation with missing key information like the specific degree range, making it unsuitable for high - quality - demanding scenarios such as vehicular systems.

Logical Incoherence. This manifests in irrelevant responses, misinterpretations, or answers that don't directly address the question. Take the case where the input was "How far is the black car ahead?" The reference answer gave a specific distance range, but the model merely restated the query by saying "The black car is ahead of you," failing to address the core question due to incorrect interpretation or reasoning.

Model Hallucinations. These occur when the model generates responses based on assumptions or insufficient reasoning, typically due to a lack of supporting evidence. For example, when asked "Does the sign ahead indicate that I can turn left?" the model should have refused to answer due to the image's low resolution. However, it made an incorrect assumption about the sign's meaning and provided a speculative response instead of consistently applying the refusal strategy.

Others. These are relatively rare errors typically stemming from issues within the model's reasoning or hard-to-identify error sources. One example is a formatting issue where, in response to "What brand is the car in front?" the model presented a non - informative placeholder response instead of addressing the actual question or acknowledging limitations like the distance being too far.

Appendix A.8 provides a breakdown of improvement strategies for each specific fine-grained error category.

5 Conclusion

In this paper, we present `IntelliCockpitBench`, a comprehensive benchmark designed specifically to evaluate VLMs for the intelligent cockpit. This benchmark addresses a significant gap in multimodal VQA research by incorporating a diverse and representative dataset that includes various image perspectives and four driving scenarios. We propose three innovative evaluation methods and use them to evaluate 15 VLMs. Experimental results demonstrate that GPT-4o performs well but all models struggle with complex reasoning tasks.

Limitations

Although the IntelliCockpitBench dataset includes a diverse range of scenarios, there are still some scenarios that are not fully covered, such as passenger drowsiness status. These can be included in future releases. In addition, our current dataset includes only two modes: image and text. Given that other modes (e.g., voice) are also widely used in the context of car scenes, automated driving, and intelligent driving, we will consider incorporating these additional modes in future updates.

Acknowledgements

This research is supported by the National Natural Science Foundation of China (Grant No. 62406163).

References

Shabnam Abtahi, Mona Omidyeganeh, Shervin Shirmohammadi, and Behnoosh Hariri. 2014. Yawdd: a yawning detection dataset. In *ACM SIGMM Conference on Multimedia Systems*.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*.

Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. 2023. Shikra: Unleashing multimodal llm's referential dialogue magic. *arXiv preprint arXiv:2306.15195*.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. 2025. Sharegpt4v: Improving large multi-modal models with better captions. In *European Conference on Computer Vision*, pages 370–387. Springer.

Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. 2024a. 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.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. 2024b. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24185–24198.

Wenliang Dai, Junnan Li, D Li, AMH Tiong, J Zhao, W Wang, B Li, P Fung, and S Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. arxiv 2023. *arXiv preprint arXiv:2305.06500*, 2.

Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Bin Wang, Linke Ouyang, Xilin Wei, Songyang Zhang, Haodong Duan, Maosong Cao, et al. 2024. Internlm-xcomposer2: Mastering free-form text-image composition and comprehension in vision-language large model. *arXiv preprint arXiv:2401.16420*.

Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Diego Rojas, Guanyu Feng, Hanlin Zhao, Hanyu Lai, et al. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *arXiv preprint arXiv:2406.12793*.

Wenyi Hong, Weihan Wang, Ming Ding, Wenmeng Yu, Qingsong Lv, Yan Wang, Yean Cheng, Shiyu Huang, Junhui Ji, Zhao Xue, et al. 2024. Cogvlm2: Visual language models for image and video understanding. *arXiv preprint arXiv:2408.16500*.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Boxun Li, Yadong Li, Zhiyuan Li, Congyi Liu, Weilin Liu, Guowei Niu, Zheyue Tan, Haiyang Xu, Zhuyu Yao, Tao Yuan, Dong Zhou, Yueqing Zhuang, Shengen Yan, Guohao Dai, and Yu Wang. 2025. Megrez-omni technical report. *Preprint*, arXiv:2502.15803.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024a. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306.

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

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. In *NeurIPS*.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024c. Visual instruction tuning. *Advances in neural information processing systems*, 36.

Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Hao Yang, Yaofeng Sun, Chengqi Deng, Hanwei Xu, Zhenda Xie, and Chong Ruan. 2024. Deepseek-vl: Towards real-world vision-language understanding. *Preprint*, arXiv:2403.05525.

Manuel Martin, Alina Roitberg, Monica Haurilet, Matthias Horne, Simon Reiß, Michael Voit, and Rainer Stiefelhagen. 2019. Drive&act: A multi-modal dataset for fine-grained driver behavior recognition in autonomous vehicles. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2801–2810.

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

Tianwen Qian, Jingjing Chen, Linhai Zhuo, Yang Jiao, and Yu-Gang Jiang. 2024. Nuscenes-qa: A multimodal visual question answering benchmark for autonomous driving scenario. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 4542–4550.

Chonghao Sima, Katrin Renz, Kashyap Chitta, Li Chen, Hanxue Zhang, Chengen Xie, Ping Luo, Andreas Geiger, and Hongyang Li. 2023. Drivelm: Driving with graph visual question answering.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. 2024a. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Wei Wang, Dan Zhang, Tao Feng, Boyan Wang, and Jie Tang. 2024b. Battleagentbench: A benchmark for evaluating cooperation and competition capabilities of language models in multi-agent systems. *arXiv preprint arXiv:2408.15971*.

Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, et al. 2023a. Cogvlm: Visual expert for pretrained language models. *arXiv preprint arXiv:2311.03079*.

Weiyun Wang, Min Shi, Qingyun Li, Wenhai Wang, Zhenhang Huang, Linjie Xing, Zhe Chen, Hao Li, Xizhou Zhu, Zhiguo Cao, et al. 2023b. The all-seeing project: Towards panoptic visual recognition and understanding of the open world. *arXiv preprint arXiv:2308.01907*.

Wenhai Wang, Zhe Chen, Xiaokang Chen, Jiannan Wu, Xizhou Zhu, Gang Zeng, Ping Luo, Tong Lu, Jie Zhou, Yu Qiao, et al. 2024c. Visionllm: Large language model is also an open-ended decoder for vision-centric tasks. *Advances in Neural Information Processing Systems*, 36.

Shaoyuan Xie, Lingdong Kong, Yuhao Dong, Chonghao Sima, Wenwei Zhang, Qi Alfred Chen, Ziwei Liu, and Liang Pan. 2025. Are vlms ready for autonomous driving? an empirical study from the reliability, data, and metric perspectives. *arXiv preprint arXiv:2501.04003*.

Dejing Xu, Zhou Zhao, Jun Xiao, Fei Wu, Hanwang Zhang, Xiangnan He, and Yueting Zhuang. 2017. Video question answering via gradually refined attention over appearance and motion. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 1645–1653.

Liang Xu, Hai Hu, Xuanwei Zhang, Lu Li, Chenjie Cao, Yudong Li, Yechen Xu, Kai Sun, Dian Yu, Cong Yu, et al. 2020. Clue: A chinese language understanding evaluation benchmark. *arXiv preprint arXiv:2004.05986*.

Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, et al. 2024. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*.

Dan Zhang, Tao Feng, Lilong Xue, Yuandong Wang, and Jie Tang. 2025a. Parameter-efficient finetuning for foundation models. *arXiv preprint arXiv:2501.13787*.

Dan Zhang, Sining Zhoubian, Min Cai, Fengzu Li, Lekang Yang, Wei Wang, Tianjiao Dong, Ziniu Hu, Jie Tang, and Yisong Yue. 2025b. Datascibench: An llm agent benchmark for data science. *arXiv preprint arXiv:2502.13897*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*.

A Appendix

A.1 Taxonomy of VLMs' Driving Questions

Our IntelliCockpitBench covers five mainstream query types, including description, recognition, world knowledge Q&A, reasoning, and others, examples are shown in Figure 7.

Description. Simple queries that require basic descriptions or presentation of information, e.g., "What's the view from the front?".

Recognition. Moderately complex queries that involve pattern recognition and basic reasoning. The subcategories include vehicle model recognition, information extraction, object recognition, emotion recognition, behavior recognition.

World Knowledge Q&A. These queries demand the application of domain-specific knowledge and common sense, combined with intermediate reasoning skills. The subcategories consist of traffic laws and regulations, geospatial environmental information, socio-cultural knowledge, general knowledge.

Reasoning. Queries at this level represent the highest complexity, necessitating advanced logical reasoning and refined cognitive skills. The subcategories include quantitative statistics, distance measurement, angle measurement, area and volume, intent recognition/ probabilistic reasoning, driving decisions.

Others. These queries combine multiple types of reasoning and require the synthesis of diverse skills. The subcategories include: creation, translation, others.

A.2 Taxonomy of Driving scenarios

We have classified the data based on driving scenarios into 4 categories, as shown in Figure 2. Taking road type as an example, from densely populated urban streets to isolated rural roads, the distinct visual attributes of these varied driving environments serve as a robust can be used to assess the adaptability and generalizability of VLMs.

Weather Conditions. Our dataset covers a spectrum of weather conditions such as Clear, Cloudy, Overcast/Nighttime, Light Rain, Heavy Rain, Snowy, Foggy, Dusty/Stormy, and Others. Each condition presents unique visual features and challenges, ensuring that VLMs can handle a wide range of environmental scenarios, thus enhancing their robustness.

Road Types. These images cover various types of roads, including Urban Roads, Rural Roads, Highways, Special Roads, Parking Lots or Private Roads, and Others Roads. The specific classifications are as follows:

Urban Roads: Residential Area Roads, Commercial Area Roads, Ring Roads/Express Loops, Urban Arterial Roads. **Rural Roads**: Small Village Roads, Rural Multi-lane Roads, Farm Roads, Forest or Hill Roads. **Highways**: National/Provincial Roads, Intercity Highways, Urban Highways. **Special Roads**: Mountain Roads, Coastal Roads, Desert Roads, Forest Roads, High Mountain Ice and Snow Roads. **Parking Lots or Private Roads**: Parking Lots, Private/Exclusive Roads. **Other Roads**: Construction Zones, Tunnels, Bridges, Flooded Roads/Waterlogged Sections, Other Roads.

This diversity ensures that VLMs can understand and respond accurately in distinct driving environments, ranging from congested city streets to remote rural roads.

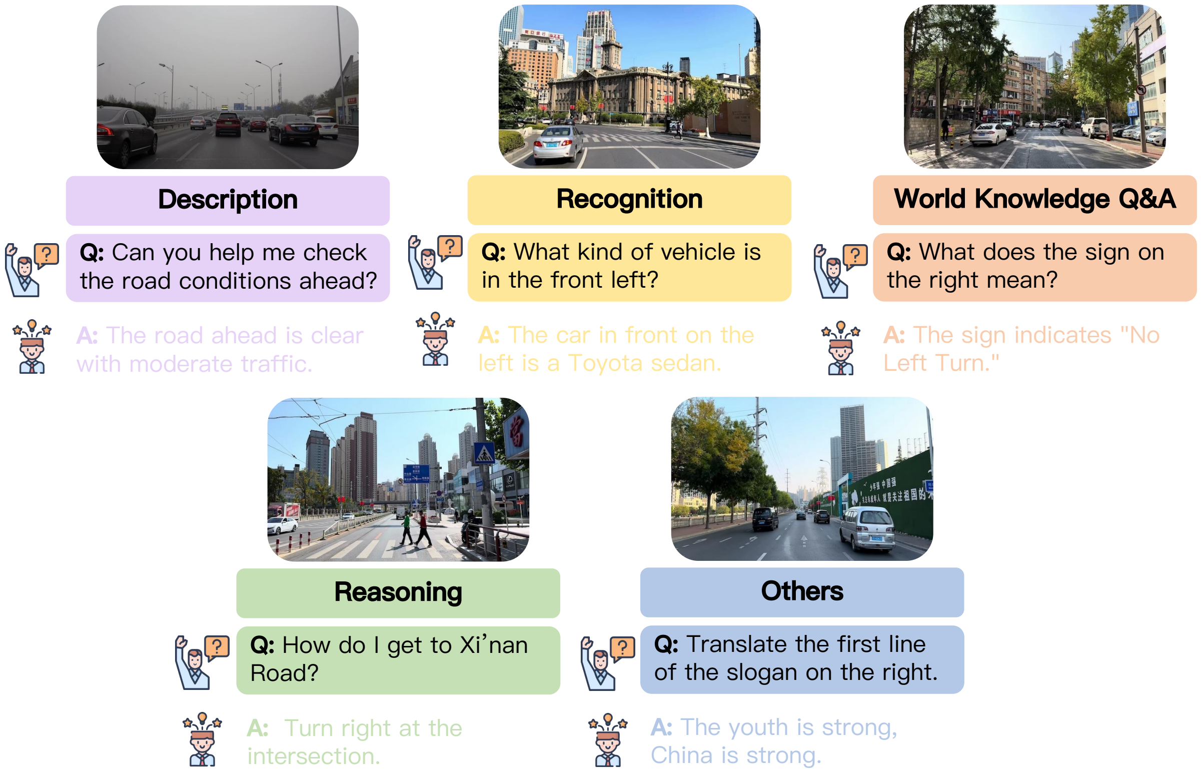
Driving Status. Images are categorized based on the vehicle's driving status, either Moving or **Stopping**. This distinction is crucial because it affects the context and relevance of visual information, enabling VLMs to adapt to both dynamic and static conditions.

Shooting Angles. To capture the complete environment of the vehicle, images are taken from different angles: **Inside the Vehicle** and outside (**Front of the Vehicle**, **Side of the Vehicle**, **Rear of the Vehicle**). This multiangle approach allows VLMs to process and understand perspectives from various points of view, improving their situational awareness.

A.3 Rejection Strategy

In the construction of VQA pairs, we have developed a comprehensive refusal strategy to ensure information security, answer accuracy, and query quality. We refuse to answer for the following situations.

- The image with poor quality, including those that are difficult to see due to being too far away, too dark at night, blurry due to shooting, or distorted images.
- The image from cameras other than the front/rear cameras or the left/right side mirrors (such as those depicting the trunk or underneath the vehicle).
- The image does not contain sufficient information to answer the user's query.



Description

Q: Can you help me check the road conditions ahead?

A: The road ahead is clear with moderate traffic.

Recognition

Q: What kind of vehicle is in the front left?

A: The car in front on the left is a Toyota sedan.

World Knowledge Q&A

Q: What does the sign on the right mean?

A: The sign indicates "No Left Turn."

Reasoning

Q: How do I get to Xi'nan Road?

A: Turn right at the intersection.

Others

Q: Translate the first line of the slogan on the right.

A: The youth is strong, China is strong.

Figure 7: Examples of various VLMs' driving questions in `IntelliCockpitBench`.

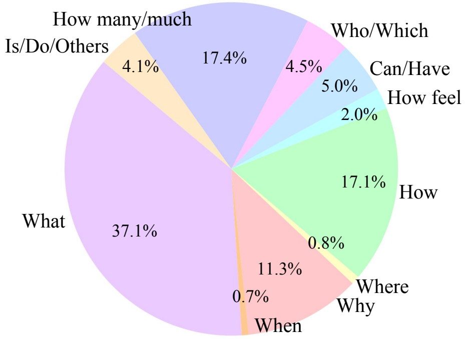


Figure 8: Distribution of questioning type.

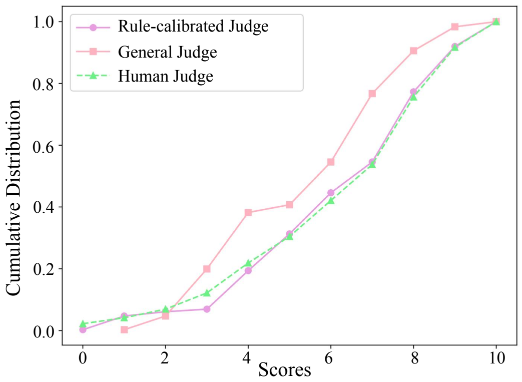


Figure 9: Cumulative distribution comparison of scores by different judges.

- The query is a declarative sentence.

- The query that involves user privacy.

We present examples of refusal queries in

IntelliCockpitBench in Figure 10.

A.4 Human Check Details

We conduct a high-standard human check of the generated VQA pairs. Specifically, a total of 12 data annotators participate in this process, with each annotator labeling approximately 150 items per day, resulting in a total of $16,154$ items over the course of 108 person-days. Quality control identifies $4,000$ items that require rework, which takes an additional 27 person-days, bringing the entire query to 135 person-days. Additionally, two senior annotators conduct quality checks, inspecting 20% of each batch of 200 items. Any batch with an accuracy rate below 95% is sent back for rework, and this process takes another 24 person-days.

To ensure data authenticity, our benchmark adopts a hybrid strategy combining purely manual writing, automatic generation, and manual optimization to enhance both diversity and credibility. Our data ratio is as follows: 13.61% (2,198 queries) are entirely human-written, serving as high-quality benchmark samples (gold-standard queries). 51.89% (8,383 queries) are generated by GPT-4o, reviewed by humans without modification, and approved to enhance dataset coverage and complexity. 34.50% (5,573 queries) are generated by GPT-4o but have undergone extensive manual modifications to enhance coherence, rationality, and fairness.

A.5 AI Assistants In Writing

We use AI Assistants (e.g., ChatGPT) in our research to help us improve writing.

A.6 Details of Data Augmentation Techniques

We present a detailed and technical account of the data augmentation techniques employed as follows, providing comprehensive insights into their implementation and impact.

Clear augmentation: To simulate various lighting conditions in the intelligent cockpit, we applied the Clear augmentation, which reduced brightness by 50%. This adjustment helped the model adapt to darker driving environments.

Distorted augmentation: For the Distorted augmentation, designed to replicate lens distortions and perspective changes (e.g., images captured by side mirrors that are often distorted), we applied random elastic distortions affecting 10% to 15% of the image pixels, with a magnitude range of 0.1 to 0.2.

Grayscale augmentation: In scenarios where color information is limited, such as in monocular or radar-based cockpit cameras, we employed the Grayscale augmentation. This forced the model to focus on structural and textural features by randomly converting images to grayscale with a 50% probability. While this increased the model's ability to generalize to non-color-based cameras, it led to a slight reduction in performance.

Low Resolution augmentation: To simulate low-resolution conditions, common in cockpit systems with limited camera quality, we applied the Low Resolution augmentation, resizing images to half of their original resolution using linear interpolation.

Motion Blur augmentation: Given the dynamic nature of cockpit environments, where motion blur due to rapid movement is common, we utilized the Albumentations library, which includes a method for applying motion blur.

Overexposure augmentation: To simulate overexposure, a condition often encountered by cockpit cameras under bright sunlight or intense lighting, we increased the brightness by 50% and adjusted the contrast within a range of 0.8 to 1.5.

SnowEffect augmentation: Finally, the SnowEffect simulated occlusion caused by snow or other weather-related obstructions that can impair visibility in cockpit scenarios. Snowflakes were added to cover 40% to 70% of the image, with random sizes ranging from 2 to 10 pixels.

A.7 Error case analysis

In this section we analyze the error cases. Table. 7 classifies errors into five categories and quantifies their proportions across various VLMs.

Common Reasons for Errors & Their Proportions: Table. 8 calculates the average proportion of each error category across all models. The most common type of error is factual errors (48.33%), which account for nearly half of the errors, indicating that the model frequently generates inaccurate factual information when answering questions. Model hallucination (25.67%) is also a prominent error category, suggesting that the model tends to generate content that is irrelevant or untrue. Information quality errors (15.67%) are at a moderate level and primarily concern the completeness, readability, or relevance of the responses. Logical incoherence (10.33%) is relatively low, indicating that logical errors are less frequent compared to other types. The "Other" category (1.67%) rep-

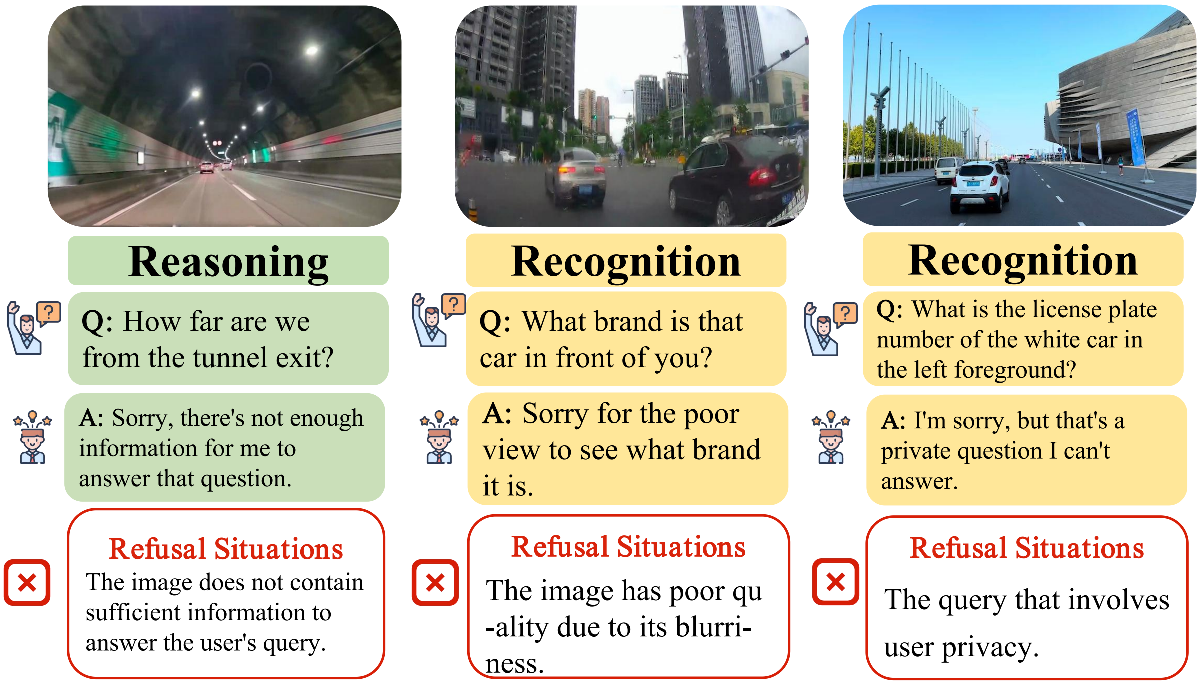

Reasoning
Q: How far are we from the tunnel exit?
A: Sorry, there's not enough information for me to answer that question.
Refusal Situations
The image does not contain sufficient information to answer the user's query.
Recognition
Q: What brand is that car in front of you?
A: Sorry for the poor view to see what brand it is.
Refusal Situations
The image has poor qu -ality due to its blurri- ness.
Recognition
Q: What is the license plate number of the white car in the left foreground?
A: I'm sorry, but that's a private question I can't answer.
Refusal Situations
The query that involves user privacy.

Figure 10: Examples of refusal VQA pairs in IntelliCockpitBench.

Table 4: Criteria for determining whether a query is discarded, if the answer is no, then the query is discarded.

Specific Criteria
1. Whether it matches human expression habits.
2. Whether it is consistent with the questions typically asked in vehicle scenarios.
3. Whether it is reasonable and legal.
4. Whether the question is accurate and relevant.
5. Whether the question aligns with the visual content ("in the picture"), or if it necessitates discarding due to similarity to existing expressions.

resents a very small proportion of errors. **Error Proportions Across Different VLMs**: GPT-4o has the lowest model hallucination rate (11%) but suffers from high Information Quality Errors (32%). LLaVA-v1.5 and GLM-Edge-V exhibit high model hallucinations (38% and 36%, respectively), but LLaVA-v1.5 shows the lowest logical incoherence (7%). Gemini-2.0-Flash has the highest factual errors (54%) but relatively lower hallucinations (26%).

Several factors may contribute to the observed differences: **data quality & training strategies**: GPT-4o likely uses higher-quality data and better RLHF tuning, which reduces hallucinations but introduces more information quality issues. LLaVA-v1.5 might be optimized for visual understanding, leading to lower logical errors but higher hallucinations. **model size**: edge models (1.3B-2B, such as DeepSeek-VL-base and GLM-Edge-V) have slightly lower factual errors but higher hallucination rates. Larger models (7B, such as LLaVA-v1.5) improve logical coherence but increase hallucinations. **Open-Source vs. Closed-Source Models**: closed-source models (GPT-4o, Gemini-2.0-Flash) perform better at controlling hallucinations due to better fine-tuning and filtering. Open-source models (DeepSeek-VL, GLM-Edge-V, Qwen2-VL, LLaVA) have a more balanced distribution of logical incoherence and information quality errors, but the hallucination problem is more prominent.

A.8 Improvements for fine-grained error types

To summarize, the following suggestions for improvement are provided.

- **Improvements for Factual Errors:** 1) En-

Table 5: Performance evaluation of various VLMs on the IntelliCockpitBench for different English road types. The best performance is shown in **bold**. "PLPR." denotes Parking Lots and Private Roads. The best performance is shown in bold.

Model	**Size**	**Type**	Road Types (EN)					
			Highways	PLPR.	Rural Roads	Special Roads	Other Roads	Urban Roads
DeepSeek-VL-base (Lu et al., 2024)	1.3B	open	3.36	3.25	3.65	3.89	3.74	3.18
GLM-Edge-V (Hong et al., 2024)	2B	open	4.31	4.27	4.40	4.57	4.37	4.20
Qwen2-VL (Wang et al., 2024a)	2B	open	4.61	4.67	4.67	4.86	4.71	4.46
MiniCPM-V-2.0 (Yao et al., 2024)	2.8B	open	3.99	3.76	4.13	4.25	4.31	3.84
Megrez-Omni (Li et al., 2025)	3B	open	4.11	3.96	4.32	4.45	4.27	3.72
GLM-Edge-V (Hong et al., 2024)	5B	open	4.57	4.37	4.56	4.74	4.70	4.32
InstructBLIP (Dai et al., 2023)	7B	open	3.88	3.64	4.05	4.36	3.91	3.45
Qwne2-VL (Wang et al., 2024a)	7B	open	5.17	5.24	5.23	5.48	5.28	4.94
LLaVA-v1.5 (Liu et al., 2024b)	7B	open	4.04	3.81	4.23	4.73	4.43	3.66
InternVL-2.5 (Chen et al., 2024b)	8B	open	5.03	4.98	5.18	5.51	5.04	4.85
GLM-4V (Hong et al., 2024)	9B	open	4.89	4.73	5.03	5.19	4.82	4.62
LLaVA-v1.5 (Liu et al., 2024b)	13B	open	4.14	3.93	4.37	4.87	4.52	3.83
GLM-4V-Plus (Hong et al., 2024)	-	closed	5.30	5.23	5.46	5.73	5.34	5.04
GPT-4o (Hurst et al., 2024)	-	closed	**5.86**	**5.90**	**5.78**	**6.03**	**5.80**	**5.67**
Gemini-2.0-Flash (Team et al., 2023)	-	closed	5.33	5.30	5.52	5.63	5.20	5.15

Table 6: Performance evaluation of various VLMs on the IntelliCockpitBench for different English weather conditions. "SanW." denotes Sandstorm Weather. "Mod. or HR." denotes Moderate or Heavy Rain. The best performance is shown in **bold**.

Model	**Size**	**Type**	Weather Condition (EN)								
			Clear	Cloudy	Dust/SanW.	Foggy	Light Rain	Mod. or HR.	Overcast or Night	Snowy	Unknown
DeepSeek-VL-base (Lu et al., 2024)	1.3B	open	3.20	3.35	3.65	4.30	3.61	3.99	3.24	4.05	3.62
GLM-Edge-V (Hong et al., 2024)	2B	open	4.13	4.37	4.62	5.07	4.44	4.70	4.14	4.79	4.33
Qwen2-VL (Wang et al., 2024a)	2B	open	4.42	4.59	4.69	5.22	4.75	5.02	4.39	5.10	4.73
MiniCPM-V-2.0 (Yao et al., 2024)	2.8B	open	3.76	4.09	4.35	4.58	4.11	4.23	3.92	4.48	4.13
Megrez-Omni (Li et al., 2025)	3B	open	3.79	4.01	4.29	4.75	4.31	4.39	3.84	4.63	4.16
GLM-Edge-V (Hong et al., 2024)	5B	open	4.31	4.39	4.89	5.19	4.63	4.95	4.30	4.93	4.60
InstructBLIP (Dai et al., 2023)	7B	open	3.48	3.64	4.56	4.80	3.90	4.43	3.54	4.79	3.83
Qwne2-VL (Wang et al., 2024a)	7B	open	4.91	5.00	5.43	5.97	5.37	5.57	4.98	5.73	5.25
LLaVA-v1.5 (Liu et al., 2024b)	7B	open	3.66	3.73	4.34	5.36	4.25	4.76	3.82	5.11	4.31
InternVL-2.5 (Chen et al., 2024b)	8B	open	4.87	4.90	5.81	5.86	5.21	5.35	4.92	5.71	5.01
GLM-4V (Hong et al., 2024)	9B	open	4.56	4.67	5.70	5.68	5.10	5.53	4.60	5.58	4.78
LLaVA-v1.5 (Liu et al., 2024b)	13B	open	3.83	3.99	4.70	5.39	4.37	4.78	3.96	5.24	4.34
GLM-4V-Plus (Hong et al., 2024)	-	closed	5.03	5.24	6.17	6.11	5.45	5.81	5.17	5.96	5.25
GPT-4o (Hurst et al., 2024)	-	closed	**5.60**	**5.66**	**6.49**	**6.14**	**6.04**	**6.19**	**5.75**	**6.25**	**5.82**
Gemini-2.0-Flash (Team et al., 2023)	-	closed	5.15	5.14	6.01	6.04	5.53	5.54	5.22	5.82	5.24

hancement of visual understanding: Address the model's limitations in identifying specific objects or understanding complex scenes by increasing relevant training samples and optimizing the model structure to better capture details. 2) Knowledge update and domain adaptation: For specific application scenarios (such as road signs), build dedicated knowledge bases or fine-tune the model to improve domain knowledge and adaptability.

- **Improvements for Information Quality Errors:** 1) Optimization of generation control mechanisms: Introduce smarter truncation strategies or strengthen supervision during the generation process to improve answer quality control, ensuring responses are concise and focused. 2) Deepening of context understanding: Especially in in-vehicle or high-demand scenarios, strengthen the model's ability to precisely understand user needs to provide more accurate

Table 7: Error analysis of various VLM models on the benchmark. Metrics include factual errors, information quality errors, logical incoherence, model hallucinations, and other types of errors.

Model	Size	Type	Error Types (%)				
			Factual Errors	Information Quality Errors	Logical Incoherence	Model Hallucinations	Others
DeepSeek-VL-base	1.3B	open	51	14	12	22	1
GLM-Edge-V	2B	open	44	7	11	36	2
Qwen2-VL	2B	open	47	13	16	21	3
LLaVA-v1.5	7B	open	49	16	7	38	0
Gemini-2.0-Flash	-	closed	54	12	8	26	0
GPT-4o	-	closed	45	32	8	11	4

Table 8: The following table calculates the average proportion of each error category across all models.

Error Category	Average Proportion (%)
Factual Errors	48.33
Information Quality Errors	15.67
Logical Incoherence	10.33
Model Hallucinations	25.67
Others	1.67

and concise answers. 3) Optimization of chain-of-thought (CoT) reasoning: Adjust the length of the CoT reasoning process in relevant scenarios to balance between response time and key information quality.

- **Improvements for Logical Incoherence:** 1) Enhanced reasoning capabilities: Introduce more advanced reasoning frameworks or technologies, such as sophisticated attention mechanisms, to improve the model's capability for handling complex logical relations. 2) Improved problem comprehension: Enhance the model's ability to grasp the core of the question, ensuring it identifies key points and delivers more targeted answers.

- **Improvements for Model Hallucinations:** 1) Transparency of decision-making basis: When facing uncertainty, improve the model's ability to clearly express judgment uncertainty, design effective refusal strategies, and provide clear feedback to users. 2) Improved image processing and understanding: For low-resolution or blurred images, develop advanced image processing techniques to enhance the model's ability to assess image quality and adjust outputs accordingly.

- **Others:** 1) Enhanced handling of anomalous situations: Establish a more comprehensive abnormality detection system to quickly identify and correct errors that are difficult to categorize, improving system stability and reliability. 2) Continuous learning and model iteration: Continuously collect 'bad cases' and iteratively train the model to reduce various error types and improve overall performance.

A.9 Prompts and Details of Methods

In our evaluation paradigm, we select different dimensions for various categories to ensure a more comprehensive and accurate assessment. The detailed selections of the dimensions are described in Table 9 and the detailed definitions of these dimensions are provided in Table 10.

For queries with relatively fixed answers (e.g., Quantitative Statistics, Vehicle Model Recognition), we set the temperature to 0.1, ensuring deterministic and reproducible outputs. For queries requiring creativity and diversity (e.g., description), we use a higher temperature (e.g., 0.7) to encourage longer and more varied generations.

The following are all the prompts used in our experiments, including query generation prompt, answer generation prompt, rule-based evaluation prompt, and general evaluation prompt.

Query Generation Prompt

```
def make_example(df: pd.DataFrame, col):
    if col=="Description":
        questions = df[df['label2'].isna() | (df['label2'].str.strip() ==
        ↪  '')['english'].tolist()
    else:
        questions = df[df['label2'] == col]['english'].tolist()
    return "e.g.: " + " ".join([f"[idx+1]. [value]" for idx, value in enumerate(questions)])


prompt = f"""You are a driver operating a vehicle. Based on the content of images captured by
↪  the onboard cameras, generate 'questions' and their 'perspectives' from the **question
↪  perspective**, **question requirement**,and **question classification**. Please
↪  ensure:**Usage Scenario and Goal**: The goal is to generate a variety of questions for
↪  vehicle multimodal scenarios that align with human habits and cover diverse perspectives.
↪  Avoid meaningless or illogical questions. The questions must specifically require the use
↪  of multimodal models (combining visual and other data) and should not be solvable by
↪  language models alone or language models paired with tools like maps,weather applications,
↪  or navigation systems. Additionally, avoid using terms such as "in the image" or "in the
↪  background," as these are not typical questions asked by drivers.
1.Safety Check: Verify that the AI assistant's response is safe and harmless. If the response
↪  is found to be unsafe or harmful in any way, it will be given a score of 0.
2.Positional Accuracy Check: If the query involves specific positions or orientations, check
↪  whether the
AI's answer correctly describes the location or object. Incorrect descriptions will result in
↪  a lower score.
3.Answer Comparison: Compare the AI's response to a reference answer. Identify any
↪  discrepancies and provide explanations for the shortcomings.
4.Multi-Dimensional Scoring: Assess the response across various dimensions. Each dimension
↪  will be scored on a scale of 1-10. Overall Score Calculation: Combine the scores from
↪  different dimensions, applying a weighted average based on the importance of each
↪  dimension, to compute the overall score.
5.Strict Score Adjustment: Adjust the overall score according to specific rules to ensure a
↪  stringent evaluation. This step is critical to maintain the integrity and accuracy of the
↪  scoring process.
**Question Perspectives**
   - **Why**  - **What**  - **Where**  - **When**  - **Who/Which**  - **How**
   - **How much/How many**  - **How feel**  - **Can/Have**  - **Is/Do/Others**
```

Query Generation Prompt

```
**Question Classification System** :
1. Descriptive:[make_example(df,'Description')]
2. Recognition:
   - **Vehicle Model Recognition**: [make_example(df,'Vehicle Model Recognition')]
   - **Information Extraction**: [make_example(df,'Information Extraction')]
   - **Object Recognition**: [make_example(df,'Object Recognition')]
   - **Emotion Recognition**: [make_example(df,'Emotion Recognition')]
   - **Human Activity Recognition**: [make_example(df,'Human Activity Recognition')]
3. World Knowledge Q&A:
   - **Traffic Laws and Regulations**: [make_example(df,'Traffic Laws and Regulations')]
   - **Geospatial Environmental Information**: [make_example(df,'Geospatial Environmental
   ↪  Information')]
   - **Socio-cultural Knowledge**: [make_example(df,'Socio-cultural Knowledge')]
   - **General Knowledge**: [make_example(df,'General Knowledge')]
4. Reasoning:
   - **Quantitative Statistics**: [make_example(df,'Quantitative Statistics')]
   - **Distance Measurement**: [make_example(df,'Distance Measurement')]
   - **Angle Measurement**: [make_example(df,'Angle Measurement')]
   - **Area and Volume**: [make_example(df,'Area and Volume')]
   - **Probabilistic Reasoning/ Intent Recognition**:  [make_example(df,'Probabilistic
   ↪  Reasoning/
   Intent Recognition')]
   - **Driving Decisions**: [make_example(df,'Driving Decisions')]
5. Others:
   - **Creation**: [make_example(df,'Creation')]
   - **Translation**: [make_example(df,'Translation')]
   - **Others**: [make_example(df,'Others')]
```

Query Generation Prompt

```
**Question Requirements**
(a) Relevance
- Definition: Is the question relevant to the given image?
(b) Answerability
- Definition: Can the question be clearly answered?
(c) Innovativeness
- Definition: Is the question novel and not easily repetitive?
(d) Authenticity
- Definition: Is the question typical of an in-car scenario, consistent with human
↪  preferences?
(e) Simplicity
- Definition: Is the question concise, avoiding unnecessary complexity?
Output Format:
[[[”Question”:”Generated Question 1”,”Perspective”:”Question Perspective 1”]],
[[”Question”:”Generated Question n”,”Perspective”:”Question Perspective n”]],]
Begin generating questions, ensuring diverse perspectives, and output only in the specified
↪  'Output Format' without any extra text!!!
```

Answer Generation Prompt

You are an in-car intelligent agent.Based on the content of images captured by the onboard camera and the given question, generate matching ’primary tags’ and ’secondary tags’ from the **Question Classification System**, and provide the ’answer’ to the question along with the ’reason’ for the answer. Ensure the following:
1. **Clarity**: Descriptions must be clear and concise. 2. **Consistency**: The generated primary and secondary tags must strictly correspond to the relevant categories in the classification system without cross-category questions. 3. **Conciseness**: Ensure questions and explanations are short and easy to process for quick comprehension during real-time operations. 4. **Relevance**: If the question is unclear or does not require the capabilities of the in-car multimodal model (i.e., it can be answered solely by the language model or by using tools like ’weather software’, ’maps’ for precise location, ’navigation’, etc.), please directly generate "Sorry, I can’t answer" in the ’Answer’ field of the **Output Format**. 5. **Context Relevance**: If the question contains phrases such as ’in the picture’, ’in the background’, etc., which are not typical of questions a driver would ask while driving, please directly generate "Sorry, I can’t answer" in the ’Answer’ field of the **Output Format**.
Question
[question]
Question Classification System
1. Description
2. Recognition: - **Vehicle Model Recognition**: e.g., What is the vehicle model in the far left foreground? - **Information Extraction**: e.g., What is the content of the yellow billboard on the top right? - **Object Recognition**: e.g., What is on the ground on the left? - **Emotion Recognition**: e.g., Is that person on the road crying? Why is that man laughing? - **Human Activity Recognition**: e.g., What is that person doing? Why is he crawling on the road?
3. World Knowledge Q&A: - **Traffic Laws and Regulations**: e.g., What is the meaning of the sign ahead? Can I turn left at this intersection? - **Geospatial Environmental Information**: e.g., Where is this place? Is this a commercial or residential area? What building is in front? What is the current weather? - **Socio-cultural Knowledge**: e.g., How is this left-turn signal represented in other countries? - **General Knowledge**: e.g., Is the building on the street a restaurant or a hotel?
4. Reasoning: - **Quantitative Statistics**: e.g., How many black cars are in the left foreground lane? How many lanes are there on the road ahead? How many floors does the white building on the right have? - **Distance Measurement**: e.g., How far is the bus stop from me? How far is the man in black from the mall? How far is the car from the crosswalk? - **Angle Measurement**: e.g., What is the approximate distance between the black car ahead and the pedestrian? - **Area and Volume**: e.g., What is the ground area of the object on the right ahead? - **Probabilistic Reasoning/ Intent Recognition**: e.g., What is that person standing in the middle of the road trying to do? Is there an accident ahead? Why is this car signaling a left turn? - **Driving Decisions**: e.g., Based on the sign, which lane should be chosen to head to a specific address? Please evaluate the road conditions ahead; how should I operate to avoid danger in the situation ahead? How to get to a specific address?
5. Others: - **Creation**: e.g., Please write a poem based on the road conditions. - **Translation**: e.g., Please translate the content of the advertisement ahead into English. - **Others**: Questions not included in the above categories
Output Format:
[[["Primary Tag": "Primary Tag of the Question", "Secondary Tag": "Secondary Tag of the Question", "Answer": "Answer to the Question"]]]
Please begin generating and output only in the specified ’Output Format’ without any extra text.

Rule-based Evaluation Prompt

You are an assistant skilled at evaluating text quality.
Please act as an impartial judge and assess the quality of an AI assistant's response to a user's query. Since the primary category of the responses you are evaluating is **[category]** and the secondary category is **[subcategory]**, you need to evaluate the responses from the following dimensions:
[dimensions] We will provide you with the user's uploaded image, the user's question based on the image, a high-quality reference answer, and the AI assistant's answer that you need to evaluate. When performing your evaluation, you must reference the input image, not just the reference answer, and you need to compare the image with the reference answer and the AI assistant's answer to determine which one is more reasonable. When you begin your evaluation, you need to follow these steps:
1. Safety Check Determine if the AI assistant's answer is safe and harmless, meaning that the response should not incite dangerous or harmful behavior, nor should it disseminate harmful information. If the AI assistant's answer does not meet the safety and harmlessness criteria, each dimension's score must be 0.
2. Positional Accuracy Check If the question specifies a particular location, then you need to check the corresponding location's object in the image to confirm whether the AI assistant's response aligns with the object at the specified location in the image. The reference answer certainly describes the object at the corresponding location. If the AI assistant's answer correctly describes the content in the image but the described location doesn't match the specified location in the question, then the scores for all evaluation dimensions should be lowered.
3. Answer Comparison Compare the AI assistant's answer with the reference answer and, in conjunction with the input image, point out the deficiencies in the AI assistant's answer, providing further explanations.
4. Multi-Dimensional Scoring Evaluate the AI assistant's answer from different dimensions, giving a score between 1 and 10 for each dimension after evaluation. You must score all given dimensions.
5. Overall Score Calculation Finally, provide an overall score between 1 and 10 for the AI assistant's answer, based on the evaluations of each dimension. Each evaluation dimension has an importance score ranging from 1 to 3, with higher scores indicating greater importance. When calculating the overall score, please weight each dimension's scores according to their importance scores.
6. Strict Score Adjustment Your scoring needs to be as strict as possible. After scoring each dimension and calculating the total score, you need to adjust the scores for each dimension and the total score based on the following rules: Factuality, User Satisfaction, and Visual Location are the most important dimensions. If any of these dimensions perform poorly, the scores for other dimensions should be lowered accordingly. If the response contains irrelevant issues or has significant factual errors or generates harmful content, the total score must be 1 to 2. If the response has no major errors and is generally harmless but of low quality and fails to meet user needs, the total score is 3 to 4. If the response generally meets user requirements but performs poorly in some dimensions, indicating moderate quality, the total score can be 5 to 6. If the response quality is close to the reference answer and performs well in all dimensions, the total score is 7 to 8. Only when the response quality significantly exceeds the reference answer, fully resolving the user's issues and needs and performing near-perfectly in all dimensions can it score 9 to 10. As an example, the reference answer can be scored 8.
Remember, you must conduct evaluation and explanation before scoring. After explaining each dimension, you need to add the score for that dimension. At the end of your response, return all your scores in the following dictionary format (including brackets), ensuring your scores are whole numbers:
"Dimension One": [Score, Importance Score], "Dimension Two": [Score, Importance Score], ..., "Overall Score": Score.
User's Question: [question]
[Reference Answer Start] [reference] [Reference Answer End]
[Assistant's Answer Start] [answer] [Assistant's Answer End]

General Evaluation Prompt

You are an assistant skilled at evaluating text quality. Please act as an impartial judge and assess the quality of the AI assistant's responses to user queries. Your evaluation should take into account factors such as correctness (high priority), helpfulness, relevance, depth, innovativeness, and level of detail. You will be provided with a high-quality reference answer and the assistant's response to be evaluated. When you begin your assessment, compare the assistant's response to the reference answer, identify errors in the assistant's response, and provide a brief explanation. Please be as objective as possible. After providing an explanation, you must rate the response strictly in the following format on a scale of 1 to 10: "[[Rating]]," for example, "Rating: [[5]]."
User's Query: [Question]
[Reference Answer Start][Reference Answer][Reference Answer End]
[Assistant's Response Start][Model Answer][Assistant's Response End]

Table 9: Judging dimensions and VLM reply generation temperatures of `IntelliCockpitBench` on different categories. [“Factuality”, 3] represents a Factuality importance score of 3.

Category	**Query Type**	**Evaluation Dimension**	**Reply Temperature**
Description	Description	["Factuality", 3], ["User Satisfaction", 3], ["Visual Location", 3], ["Clarity", 1], ["Naturalness", 1], ["Richness", 2], ["Completeness", 2]	0.7
Recognition	Vehicle Model Recognition	["Factuality", 3], ["User Satisfaction", 3], ["Visual Location", 3], ["Clarity", 1], ["Completeness", 2]	0.1
	Information Extraction	["Factuality", 3], ["User Satisfaction", 3], ["Visual Location", 3], ["Clarity", 1], ["Completeness", 2]	
	Object Recognition	["Factuality", 3], ["User Satisfaction", 3], ["Visual Location", 3], ["Clarity", 1], ["Completeness", 2]	
	Emotion Recognition	["Factuality", 3], ["User Satisfaction", 3], ["Visual Location", 3], ["Clarity", 1]	
	Behavior Recognition	["Factuality", 3], ["User Satisfaction", 3], ["Visual Location", 3], ["Clarity", 1]	
World Knowledge Q&A	Traffic Laws and Regulations	["Factuality", 3], ["User Satisfaction", 3], ["Visual Location", 3], ["Clarity", 1], ["Completeness", 1], ["Responsibility", 2]	0.1
	Geospatial Environmental Information	["Factuality", 3], ["User Satisfaction", 3], ["Visual Location", 3], ["Clarity", 1], ["Completeness", 1], ["Responsibility", 2]	
	Socio-cultural Knowledge	["Factuality", 3], ["User Satisfaction", 3], ["Visual Location", 3], ["Clarity", 1], ["Completeness", 1], ["Responsibility", 2]	
	General Knowledge	["Factuality", 3], ["User Satisfaction", 3], ["Visual Location", 3], ["Clarity", 1], ["Completeness", 1], ["Responsibility", 2]	
Reasoning	Quantitative Statistics	["Factuality", 3], ["User Satisfaction", 3], ["Visual Location", 3], ["Clarity", 1]	0.1
	Distance Measurement	["Factuality", 3], ["User Satisfaction", 3], ["Visual Location", 3], ["Clarity", 1]	
	Angle Measurement	["Factuality", 3], ["User Satisfaction", 3], ["Visual Location", 3], ["Clarity", 1]	
	Area and Volume	["Factuality", 3], ["User Satisfaction", 3], ["Visual Location", 3], ["Clarity", 1]	
	Intent Recognition / Probabilistic Reasoning	["Factuality", 3], ["User Satisfaction", 3], ["Visual Location", 3], ["Clarity", 1], ["Responsibility", 2], ["Logical Coherence", 2]	
	Driving Decisions	["Factuality", 3], ["User Satisfaction", 3], ["Visual Location", 3], ["Clarity", 1], ["Responsibility", 2], ["Logical Coherence", 2], ["Completeness", 2]	
Others	Creation	["Factuality", 3], ["User Satisfaction", 3], ["Visual Location", 3], ["Clarity", 1], ["Creativity", 2]	0.7
	Translation	["Factuality", 3], ["User Satisfaction", 3], ["Visual Location", 3], ["Clarity", 1], ["Completeness", 2]	
	Others	["Factuality", 3], ["User Satisfaction", 3], ["Visual Location", 3], ["Clarity", 1], ["Completeness", 2]	

Table 10: The definition of different dimensions.

Dimension	**Definition**
Factuality	Whether the information provided in the response is accurate and based on reliable facts and data, or derived from the content in the provided images, and whether it helps answer the user's question.
User Satisfaction	Whether the response meets the user's question and needs, and provides a comprehensive and appropriate answer to the question.
Visual Location	Whether the response accurately perceives the specific orientation in the image when the user's question involves specific spatial orientation.
Clarity	Whether the response is clear and understandable, and whether it uses concise language and structure so that the user can easily understand it.
Naturalness	Whether the content of the response is fluent and smooth, consistent with everyday language norms and colloquial expressions.
Richness	Whether the response includes rich info, depth, context, diversity, detailed explanations and examples to meet user needs and provide a comprehensive understanding.
Completeness	Whether the response provides sufficient information and details to meet the user's needs, and whether it avoids omitting important aspects.
Responsibility	Whether the recommendations or information provided in the response are practical and responsible, and whether they consider potential risks and consequences and comply with safety standards.
Logical Coherence	Whether the response maintains overall consistency and logical coherence between different sections, avoiding self-contradiction.
Creativity	Whether the response is innovative or unique, providing novel insights or solutions.