

Are Vision LLMs Road-Ready? A Comprehensive Benchmark for Safety-Critical Driving Video Understanding

Tong Zeng
Computer Science
Virginia Polytechnic Institute and
State University
Blacksburg, Virginia, USA
tongzeng@vt.edu

Longfeng Wu
Computer Science
Virginia Polytechnic Institute and
State University
Blacksburg, USA
longfengwu@vt.edu

Liang Shi
Statistics
Virginia Tech Transportation Institute
Virginia Polytechnic Institute and
State University
Blacksburg, USA
sliang@vt.edu

Dawei Zhou
Computer Science
Virginia Polytechnic Institute and
State University
Blacksburg, USA
zhoud@vt.edu

Feng Guo*
Statistics
Virginia Tech Transportation Institute
Virginia Polytechnic Institute and
State University
Blacksburg, USA
feguo@vt.edu

Abstract

Vision Large Language Models (VLLMs) have demonstrated impressive capabilities in general visual tasks such as image captioning and visual question answering. However, their effectiveness in specialized, safety-critical domains like autonomous driving remains largely unexplored. Autonomous driving systems require sophisticated scene understanding in complex environments, yet existing multimodal benchmarks primarily focus on normal driving conditions, failing to adequately assess VLLMs' performance in safety-critical scenarios. To address this, we introduce **DVBench**—a pioneering benchmark designed to evaluate the performance of VLLMs in understanding safety-critical driving videos. Built around a hierarchical ability taxonomy that aligns with widely adopted frameworks for describing driving scenarios used in assessing highly automated driving systems, DVBench features 10,000 multiple-choice questions with human-annotated ground-truth answers, enabling a comprehensive evaluation of VLLMs' capabilities in perception and reasoning. Experiments on 14 state-of-the-art VLLMs, ranging from 0.5B to 72B parameters, reveal significant performance gaps, with no model achieving over 40% accuracy, highlighting critical limitations in understanding complex driving scenarios. To probe adaptability, we fine-tuned selected models using domain-specific data from DVBench, achieving accuracy gains ranging from 5.24 to 10.94 percentage points, with relative improvements of up to 43.59%. This improvement underscores the

necessity of targeted adaptation to bridge the gap between general-purpose vision-language models and mission-critical driving applications. DVBench establishes an essential evaluation framework and research roadmap for developing VLLMs that meet the safety and robustness requirements for real-world autonomous systems. We released the benchmark toolbox and the fine-tuned model at: <https://github.com/tong-zeng/DVBench.git>.

CCS Concepts

• **Computing methodologies** → **Computer vision; Scene understanding; Natural language processing; Neural networks.**

Keywords

DVBench; Benchmark; Driving Scene Understanding; Scene Understanding; Video Understanding; Vision Large Language Models

ACM Reference Format:

Tong Zeng, Longfeng Wu, Liang Shi, Dawei Zhou, and Feng Guo. 2025. Are Vision LLMs Road-Ready? A Comprehensive Benchmark for Safety-Critical Driving Video Understanding. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2 (KDD '25)*, August 3–7, 2025, Toronto, ON, Canada. ACM, New York, NY, USA, 12 pages. <https://doi.org/10.1145/3711896.3737396>

1 Introduction

Vision Large Language Models (VLLMs) [1, 10, 18, 34, 53] have significantly advanced artificial intelligence by enabling machines to interpret and generate multimodal content, effectively bridging visual and textual data. Although VLLMs have demonstrated impressive performance in general-purpose applications, their potential in specialized, safety-critical domains remains underexplored. For example, Autonomous driving systems (ADS) require a comprehensive understanding of complex, dynamic environments, demanding accurate perception and interpretation to make safe, reliable decisions. This naturally raises the question: *Are existing VLLMs ready for on-road deployment, especially in safety-critical scenarios?*

*Corresponding Author: feguo@vt.edu



This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

KDD '25, Toronto, ON, Canada

© 2025 Copyright held by the owner/author(s).

ACM ISBN 979-8-4007-1454-2/2025/08

<https://doi.org/10.1145/3711896.3737396>

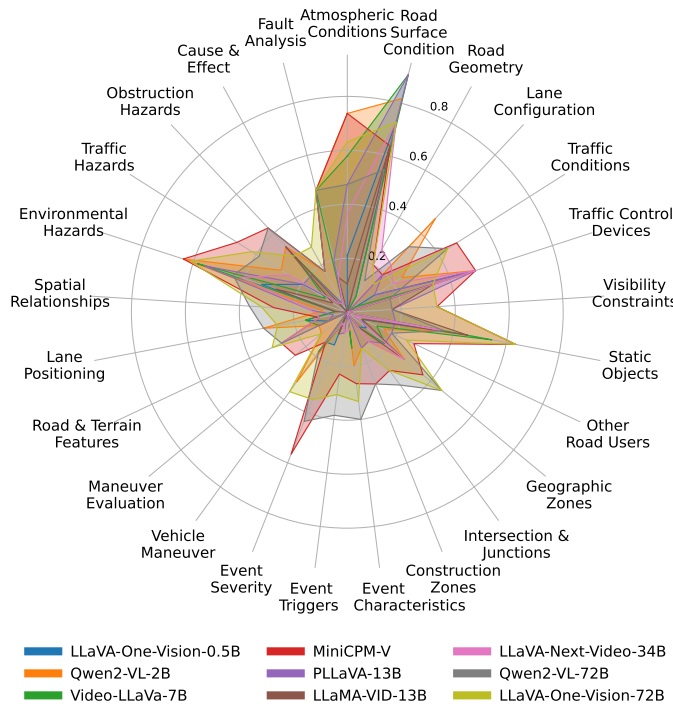


Figure 1: Performance of nine representative Vision Large Language Models, ranging from 0.5B to 72B parameters, evaluated across the 25 driving video comprehension abilities.

To answer this question, we investigate whether driving understanding generated by VLLMs reflects genuine perceptions of the environment and logical reasoning derived directly from video cues. Recent work [8, 24, 36, 46] highlights the promise of VLLMs for ADS scene understanding. However, to ensure their effective performance in these demanding contexts, it is essential to develop comprehensive benchmarks that rigorously assess their capabilities. Evaluation remains nontrivial due to the unique challenges of ADS. First, **(C1. Rarity of Safety-Critical Events)** safety-critical incidents are usually rare compared to routine driving conditions, resulting in highly imbalanced datasets. Many existing benchmarks—built on popular driving datasets such as NuScenes [27], BDD [13], and Waymo Open [32] are dominated by normal driving scenarios. This imbalance can inadvertently reward overly simplistic responses (e.g., "Going Ahead") with high accuracy [43], even though safety-critical driving scenarios, which are of paramount importance for ADS, receive insufficient attention. Second, **(C2. Lack of Temporal-Spatial Understanding in Driving Contexts)** existing benchmarks often focus on textual information or static, single-frame images, limiting their ability to capture dynamic driving scenarios. For instance, NuScenesQA [27] primarily relies on simple, one-word answers, while DriveMLLM [7] pairs front-facing camera images with linguistically diverse natural language questions. However, they largely overlook temporal-spatial perception and reasoning, essential for tracking changes over time—such as monitoring traffic signals or recognizing events leading to a crash. As a result, they fail to fully assess a model’s ability to perceive, reason, and predict in safety-critical situations. Although general-purpose multimodal benchmarks [5, 15, 20] incorporate temporal

and spatial reasoning, they are not designed for safety-critical driving scenarios. They lack evaluations specific to autonomous driving, such as interpreting traffic signs under varying conditions, analyzing pedestrian behavior, or making split-second decisions in dynamic environments. Last, **(C3. Narrow Evaluation Scope)** some datasets prioritize specific aspects of driving. For example, DRAMA [23] focuses on joint risk localization and captioning, while Rank2Tell [28] emphasizes important object identification. Moreover, the construction of questions often relies on predefined templates or limited contexts, further narrowing the scope of evaluation. Consequently, these benchmarks often lack a comprehensive understanding of driving knowledge, leading to unstable and unreliable performance in real-world applications.

To overcome these limitations, we introduce DVBenCh, a comprehensive benchmark designed to holistically evaluate VLLMs across a range of driving-specific capabilities (e.g., safety-critical events), including nuanced perception and reasoning. In particular, to address C1, the questions of DVBenCh are curated from videos of safety-critical events, such as crashes and near-crash incidents, ensuring that the evaluation covers a realistic and diverse range of challenging scenarios. To address C2, the DVBenCh questions integrate both textual descriptions and corresponding video data, ensuring a robust evaluation of VLLMs’ ability to interpret and integrate multimodal information effectively. This dual-input approach enables a more nuanced assessment of models’ ability to handle complex, safety-critical scenarios. Finally, to address C3, DVBenCh is constructed around a hierarchical ability taxonomy that aligns with the requirements for scenario-based testing of ADS established by PEGASUS (Project for the Establishment of Generally Accepted quality criteria, tools and methods as well as Scenarios and Situations) [2] and driving scenario framework by the National Highway Traffic Safety Administration (NHTSA) [35]. This taxonomy includes a three-level hierarchy, with 2 level-one (L1) abilities, 10 level-two (L2) abilities, and 25 level-three (L3) abilities, covering a spectrum of driving video understanding tasks from perception to reasoning. Grounding the taxonomy in high-level ADS requirements ensures that evaluated capabilities are directly relevant to real-world applications. Additionally, DVBenCh includes an automated annotation framework to generate multiple-choice questions for each task. This framework supports a curated dataset with diverse scenarios enriched by human annotations to ensure quality and reliability. To enhance the robustness of evaluations, we introduce GroupEval, a novel assessment strategy that analyzes models based on grouped capabilities and incorporates randomized option ordering. This approach mitigates position-based biases observed in large language models [12, 19], leading to fairer and more accurate assessments.

Our extensive experiments evaluate 14 state-of-the-art VLLMs of varying architectures and sizes, both with and without domain-specific knowledge. Results reveal significant performance gaps—no model exceeding 40% accuracy under our GroupEval strategy, underscoring critical limitations in high-level driving perception and reasoning (see Figure 1). To mitigate these gaps, we fine-tuned selected models with DVBenCh data, achieving notable performance gains. These findings underscore the need for targeted adaptation and specialized domain knowledge to bridge the gap between general-purpose models and safety-critical driving applications.

DVBench serves as a valuable benchmark for advancing VLLMs in autonomous driving by identifying model weaknesses and providing a structured evaluation framework to improve traffic safety. The main contributions of this paper are summarized as follows:

- (1) **Problem Identification:** We are among the first to investigate VLLMs’ capabilities in perception and reasoning within safety-critical (Crash, Near-Crash) driving scenarios and systematically define the hierarchical abilities for evaluating the safety of autonomous driving systems in high-risk contexts.
- (2) **Benchmark Development:** We introduce DVBench, the first comprehensive benchmark for safety-critical driving video understanding, featuring 10,000 curated multiple-choice questions across 25 key driving-related abilities. DVBench is designed to rigorously assess perception and reasoning in dynamic driving environments.
- (3) **Systematic Evaluation:** We evaluate 14 state-of-the-art VLLMs, providing an in-depth analysis of their strengths and limitations in safety-critical driving scenarios. This paper establishes structured evaluation protocols and infrastructure to enable fair comparisons and guide future advancements in VLLMs for autonomous driving.
- (4) **Open-Source Resources:** To ensure accessibility and reproducibility, we publicly release the DVBench evaluation toolbox and fine-tuned models at: <https://github.com/tong-zeng/DVBench.git>.

2 Related Work

VLLMs for Driving Scene Understanding. The application of VLLMs in ADS is an emerging area of research. DriveVLM combines VLLMs with traditional driving pipelines to address spatial reasoning challenges [36]. DriveScenify leverages VLLMs to generate context-aware responses from driving videos, aiming to enhance road safety [6]. Shoman et al. [30] propose a system that integrates object detection, tracking, and language generation for detailed traffic event descriptions. Jain et al. [11] enhance traffic understanding by incorporating VLLMs with multi-sensor data. DriveGPT-4 introduces an end-to-end ADS that utilizes multimodal large language models for video input processing and natural language interaction [46]. DME-Driver employs a VLLM with precise 3D scene perception to improve decision-making and control by integrating human decision logic into an ADS [8]. GPT-Driver repurposes the OpenAI GPT-3.5 model as a motion planner for ADS by framing motion planning as a language modeling task [24]. ScVLM integrates supervised and contrastive learning to enhance the understanding and description of safety-critical events [29]. While these studies demonstrate unique advantages in individual sub-modules such as perception, prediction, and planning, they lack comprehensive quantitative evaluations to fully assess their overall performance.

VLLM Evaluation Benchmarks. Benchmarking plays a critical role in evaluating the performance of multimodal models and guiding future research directions. Popular benchmarks such as MMBench [20], MV-Bench [15], and VideoMME [5] are designed to assess multimodal models across diverse tasks using structured evaluation methodologies, including multiple-choice questions and pre-defined tasks, ensuring fairness and consistency in scoring. While these benchmarks provide a general framework for multimodal

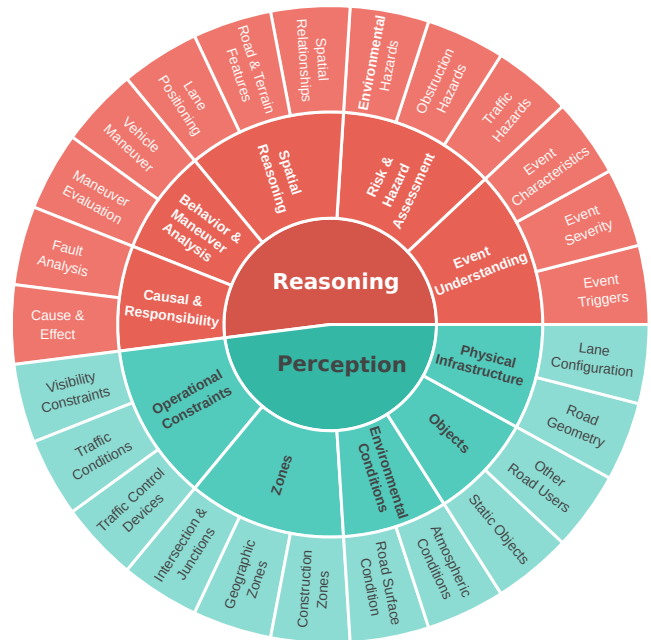


Figure 2: The DVbench Ability Hierarchy is structured across three distinct levels of capability dimensions: 2 foundational L1 abilities (perception and reasoning), 10 specialized L2 abilities for key cognitive tasks, and 25 granular L3 abilities that capture specific assessment criteria.

model evaluation, they fail to address the unique challenges inherent in autonomous driving environments. Several driving-specific VLLM benchmarks have been proposed. Reason2Drive [26] emphasizes interpretable reasoning in complex driving environments. NuScenes-QA [27] provides 34,000 multi-modal visual scenes and 460,000 question-answer pairs derived from street-view data, including images and LiDAR point clouds, to assess reasoning capabilities in challenging driving scenarios. LingoQA [25] incorporates both action and scenery datasets, but its free-form answers (averaging 17.2 words) introduce significant uncertainty in evaluation outcomes. While these benchmarks address specific tasks in autonomous driving video-language learning models (VLLMs), they do not evaluate comprehensively and fail to adequately assess performance in safety-critical driving scenarios.

3 The Proposed DVbench

In this section, we first present a hierarchical ability taxonomy organizing essential perception and reasoning skills, and then describe data collection and human annotation processes.

3.1 Hierarchical Ability Taxonomy

Driving scenarios encompass complex environments, road infrastructure, traffic, vehicles, and driver information. Established frameworks like PEGASUS [2] and NHTSA [35], define a six-layer structure for describing driving scenes, covering road infrastructure, temporary changes, objects, environment, and digital information. This layered taxonomy aligns with human perception and reasoning, categorizing L1 into "Scenario Perception" and "Scenario Reasoning", which branch into six L2 scenario layers. Integrating

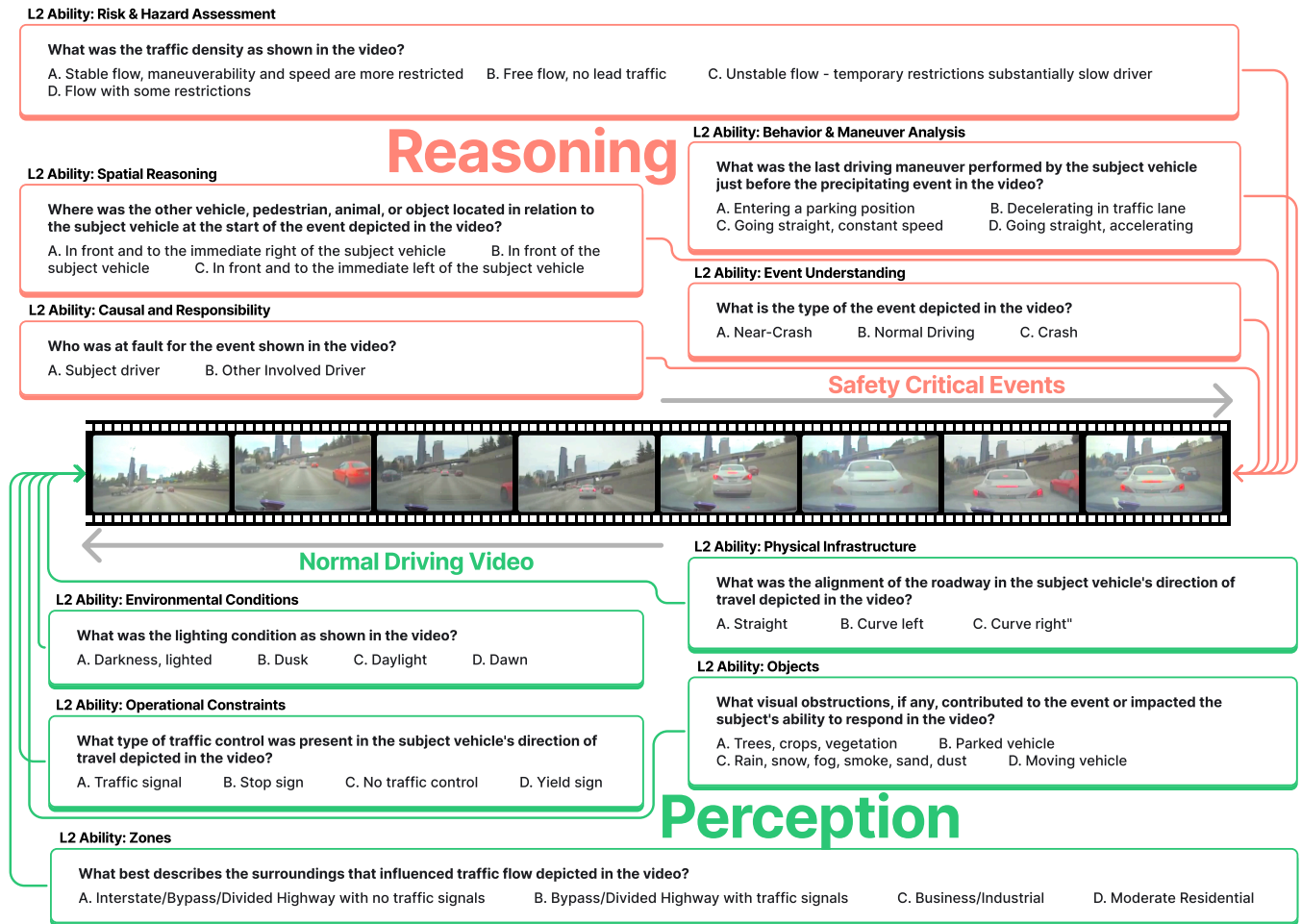


Figure 3: Examples of L2 Ability Taxonomy questions: perception and reasoning. Perception abilities focus on low-level skills such as object recognition, environmental condition parsing, and infrastructure comprehension, while reasoning abilities encompass higher-level cognitive capacities like inference, prediction, and causal analysis.

this taxonomy with driving automation levels offers a structured approach to evaluating ADS across the full spectrum of traffic safety considerations. Scenario Perception focuses on gathering information about the physical road and traffic infrastructure, including temporary changes. Scenario Reasoning then analyzes dynamic objects, environmental conditions, and digital data to assess the full context. This hierarchical taxonomy offers a structured framework for evaluating autonomous vehicle safety across diverse traffic scenarios. Figure 2 presents the ability taxonomy for evaluating VLLMs at varying abstraction levels, Figure 3 gives examples of L2 ability, and definitions are provided in Appendix A.

The video data in DV Bench were manually collected from the Second Strategic Highway Research Program (SHRP 2) [9], the largest study of naturalistic driving behaviors to date. Over 3,300 vehicles from six sites in the US were instrumented with a data acquisition system that collected four video views (driver’s face, driver’s hands, forward roadway, rear roadway), vehicle network information (e.g., speed, brake, accelerator position), and information from additional sensors included with the data acquisition system (e.g., forward radar, accelerometers). The driving data were collected

at 10 or 15 Hz. The final SHRP2 NDS data collection consists of 4,300 person-years of naturalistic driving data between 2010 and 2013 with more than one million hours of continuous driving data. A sophisticated multi-stage process was conducted in SHRP2 NDS [9] to identify all safety-critical events (SCEs), including crashes and near-crashes. The videos of SCEs were manually analyzed and annotated through a rigorous data reduction protocol [9]. The annotation includes 75 variables spanning three categories: Event Variables, Driver Variables, and Environmental Variables.

To ensure quality in data annotation, the SHRP 2 project employs a structured four-phase quality assurance and control workflow: protocol development, annotator training, active data reduction with ongoing checks, and final verification for consistency and coherence. This rigorous process maintains high data quality throughout the annotation life cycle. The annotated naturalistic driving data enriches crash analysis with detailed real-world insights.

3.2 Multiple-Choice Question Bank Creation

After obtaining the Normal Driving and Safety Critical Event videos with their annotations, we built a Multiple-Choice Question (MCQ)

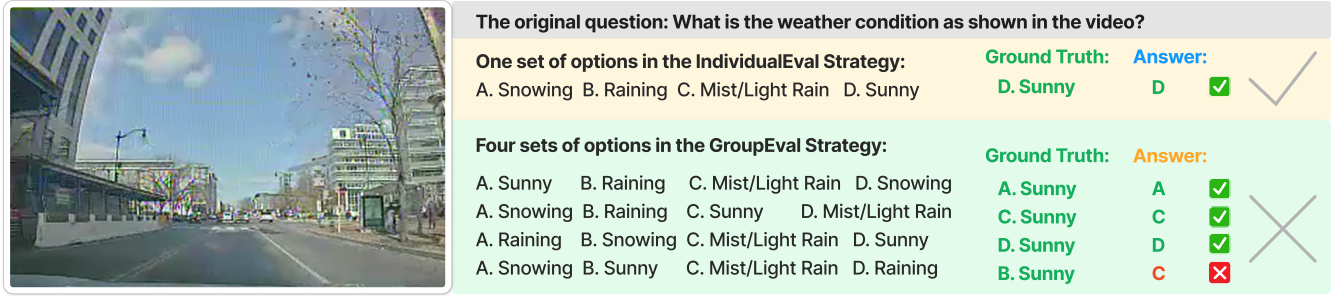


Figure 4: GroupEval Strategy. In GroupEval, each question is tested multiple times, with the correct answer’s position changing each time while the other options are shuffled randomly. Here, the IndividualEval strategy deems the VLLM successful, whereas GroupEval considers it unsuccessful, as the VLLM fails to consistently identify the correct answer across different trials.

Table 1: Comparison of our dataset with previous vision-language datasets related to autonomous driving (AD). "QA" stands for "Question and Answer," and "MCQ" denotes "Multiple Choice Question." A "-" indicates publicly unavailable data.

Benchmark	Tasks		Driving Scenario			Data Type			Modality			Amount
	Perception	Reasoning	Normal	Crash	Near-Crash	Caption	QA	MCQ	Text	Image	Video	
BDD-X [13]	✓	✗	✓	✗	✗	✓	✗	✗	✓	✓	✓	26k
BDD-OIA [45]	✓	✗	✓	✗	✗	✓	✗	✗	✓	✓	✓	35,366
Talk2car [4]	✓	✗	✓	✗	✗	✓	✗	✗	✓	✓	✗	2,447
DRAMA [23]	✓	✗	✓	✗	✗	✗	✓	✗	✓	✓	✓	17,786
NuScenesQA[27]	✓	✗	✓	✗	✗	✗	✓	✗	✓	✓	✗	83,337
Rank2Tell [28]	✓	✗	✓	✗	✗	✗	✓	✗	✓	✓	✓	-
DriveMLLM [7]	✓	✗	✓	✗	✗	✗	✓	✗	✓	✓	✗	4,666
SUP-AD [36]	✓	✗	✓	✗	✗	✓	✗	✗	✓	✓	✗	-
DriveLM [31]	✓	✓	✓	✗	✗	✓	✗	✗	✓	✓	✗	15,480
IDKB [21]	✓	✗	✓	✗	✗	✗	✓	✓	✓	✓	✗	1M
DriveBench [43]	✓	✓	✓	✗	✗	✗	✓	✓	✓	✓	✗	20,498
DVBench (Ours)	✓	✓	✓	✓	✓	✗	✗	✓	✓	✓	✓	10,000

bank using 75 expert-annotated variables per video. This MCQ bank is designed to assess the perception and reasoning abilities of VLLMs in understanding driving scenarios. We have curated a total of about 10,000 (10K) MCQs with human-annotated ground truth answers spanning 25 distinct L3 abilities. A comparison between DVBench and existing benchmarks is summarized in Table 1.

Each vision-language question-answer is structured as a multiple-choice problem, defined as a quintuple $(Q_i, C_i, V_i, A_i, K_i)$, where Q_i is the question, C_i is a set of n answer options (c_1, c_2, \dots, c_n) with $2 \leq n \leq 4$, V_i is the associated video, A_i is the ground truth answer, and K_i provides domain knowledge, including explanations of driving terms and test questions. In DVBench, each question-choice-answer set corresponds to a single video. The construction details for each element of the quintuple are introduced below.

Video Selection and Processing. DVBench consists of 4,000 video clips with a 5:4:1 ratio for Normal Driving, Crash, and Near-Crash events, extracted from 2,000 safety-critical videos. To capture short-term visual information for autonomous driving, we extracted a critical 5-second segment from each video, following the annotated event point for Crash and Near-Crash cases. The preprocessing steps include cropping timestamp watermarks and black borders, standardizing resolution and aspect ratio, resulting in 432x324 clips with a 4:3 aspect ratio and a 5-second duration.

LLM-Assisted Question Rewriting. The original video data contained 75 attributes essential for understanding and evaluating

autonomous driving scenarios. Using these annotations, we initially crafted a question set based on data guidelines. To better integrate domain knowledge, improve question-answer alignment, and ensure consistency, we drew inspiration from prior works [15, 54]. In DVBench, we used prompt templates with ChatGPT o1-preview and Qwen2.5-72B models to refine questions. The process involved: 1) Filling a predefined template with manually crafted questions, annotation guidelines, candidate answers, and correct answers. 2) Feeding the prompt into the LLM to generate alternatives. 3) Manually selecting the best outputs for the question pool.

Ground Truth and Choice Selection. To generate multiple-choice questions, we used annotated attributes as ground truth answers and drew candidate choices from other video annotations. Question quality was ensured through a dual-filtering process: removing choices that were either frequently or rarely selected, as well as questions that all LLMs consistently answered either correctly or incorrectly. Additionally, for the multiple-choice options, we selected the $n - 1$ choices from the candidate pool that were closest in length to the ground truth, thereby minimizing potential biases related to option length in model performance.

Domain Knowledge Generation. To improve the VLLM’s ability to accurately interpret and respond to the questions, we provided supplementary explanations encompassing traffic and driving-related terminology and context-specific scenarios. These

explanations were carefully crafted based on the attribute annotation guidelines to reinforce understanding of key concepts critical for autonomous driving. We refer to this supplementary information as domain knowledge intended to bridge potential gaps in the VLLM’s comprehension of specialized driving contexts. By integrating domain knowledge, we aim to ensure that VLLM can more effectively handle the unique challenges posed by driving scenarios, ultimately enhancing the reliability and relevance of its responses.

3.3 Quality Control

To ensure relevance, accuracy, and consistency, we employ two key quality control mechanisms during data construction.

Manual Checking. We conduct manual checking for the results of LLM-assisted question rewriting and domain knowledge generation to ensure that the outputs generated by the LLM align with the expectation and to avoid LLM hallucinations. Additionally, we perform multiple rounds of sampling-checking-correction iterations for the generation of question-answer pairs. Manual cross-checking also helps identify ambiguities or inconsistencies that may arise during automated extraction, enabling necessary refinements and adjustments to enhance overall data quality.

VLLM Majority Voting. To further enhance question quality, we employ a majority voting strategy involving multiple state-of-the-art VLLMs. Specifically, several models generate responses for each sampled question, and those consistently yielding unanimous incorrect (or correct) answers are flagged as potentially misleading or overly simplistic. This consensus-based quality check filters out low-quality or ambiguous questions, thereby significantly improving the robustness and reliability of the benchmark dataset.

Through quality control mechanisms, we uncovered several issues impacting the evaluation process: **1) Lack of Visual Clarity.** Certain answers cannot be directly inferred from the video. For instance, lane counts such as "6-lane" or "7-lane" are often unclear, while conditions such as "Snowy" or "Dry" are relatively easier to identify compared to less obvious labels like "Icy." **2) Skewed Answer Distribution:** Some questions exhibit imbalanced answer distributions. For example, as most vehicles in the dataset did not roll over, the predominant answer to this question is "No," significantly reducing its discriminative power. **3) Extreme Difficulty Levels.** Questions that were consistently answered incorrectly or correctly by all VLLMs were excluded, as they lack the variability required for effective benchmarking. **4) Similar Answer Options.** Subtly different choices, such as "Rear-end, striking" vs. "Rear-end, struck," or "Parking lot entrance/exit" vs. "Parking lot, within the boundary," is too challenging for current VLLMs to distinguish reliably. **5) Ambiguous Terms.** Certain Terms, such as "roadway grade," referring to road gradient or steepness, can cause confusion without further contextual clarification. **6) Mixed Answer Formats.** Questions that combine numerical and textual options pose additional interpretation challenges for VLLMs, complicating their ability to interpret and select the correct answer.

4 Experiments

In this section, we evaluate a range of VLLMs on DVBench using different evaluation strategies. Additionally, we examine the impact of fine-tuning these models with DVBench data.

4.1 Experimental Settings

We evaluate various VLLMs that capable of video understanding on DVBench, including LLaMA-VID [16], LLaVA-Next-Video [50], LLaVA-One-Vision [14], MiniCPM-V (2.6) [47], PLLaVA [44], Video-ChatGPT [22], Video-LLaVa [17], Qwen2-VL [39]. For a fair comparison, we adopt the zero-shot setting to infer DVBench questions with all VLLMs based on the same prompt template (Please refer to the Appendix for the prompt details). In the Appendix, we provide detailed information regarding the architecture and the parameter size for all Open-Source VLLMs evaluated in this paper, as well as additional results for more VLLMs under various settings.

4.2 Evaluation Strategy

DVBench employs a multiple-choice question format, which introduces inherent evaluation challenges. Random guessing could yield approximately 25% top-1 accuracy for 4-choice questions, potentially diminishing the observable performance differences among VLLMs. Furthermore, systematic biases in VLLMs complicate accurate performance assessment. For instance, many VLLMs exhibit position bias, disproportionately favoring specific answer positions, such as always selecting the first or last option [3, 19]. This tendency is further corroborated by Figure 5, highlighting the complexities of evaluating VLLMs fairly and reliably.

To address these challenges, we propose Group Evaluation (GroupEval), a more robust strategy. Each question is presented to VLLMs N times (where N equals the number of answer choices), with correct answer’s position rotates systematically while other options are randomly shuffled (as illustrated in Figure 4). A VLLM passes a question only if it identifies the correct answer in all N trials. This approach improves reliability while minimizing computational cost by terminating evaluation early if an incorrect prediction occurs.

Table 2: Comparison of IndividualEval and GroupEval strategies on DVBench. VLLM accuracy is reported for both strategies, with Δ denotes the drop when switching to GroupEval.

VLLMs	Ind. Eval	Grp. Eval	Δ
LLaMA-VID-7B [16]	24.9%	6.4%	- 18.5%
LLaMA-VID-13B [16]	18.8%	8.1%	- 10.7%
LLaVA-One-Vision-0.5B [14]	31.6%	8.1%	- 23.5%
Qwen2-VL-7B [39]	36.2%	10.6%	- 25.6%
LLaVA-Next-Video-7B [50]	33.0%	15.5%	- 17.5%
Video-LLaVa-7B [17]	37.6%	18.0%	- 19.6%
PLLaVA-7B [44]	34.9%	17.3%	- 17.6%
LLaVA-Next-Video-34B [50]	33.1%	19.8%	- 13.3%
LLaVA-One-Vision-7B [14]	38.6%	21.2%	- 17.4%
PLLaVA-13B [44]	35.2%	21.6%	- 13.6%
Qwen2-VL-72B [39]	42.1%	28.6%	- 13.5%
Qwen2-VL-2B [39]	42.8%	26.9%	- 15.9%
MiniCPM-V [47]	46.0%	29.7%	- 16.3%
LLaVA-One-Vision-72B [14]	45.0%	35.7%	- 9.3%

4.3 IndividualEval vs GroupEval Results

In this subsection, we compare two evaluation strategies: GroupEval (requiring consistent, correct predictions across multiple trials) and IndividualEval (single-trial inference). Table 2 shows results

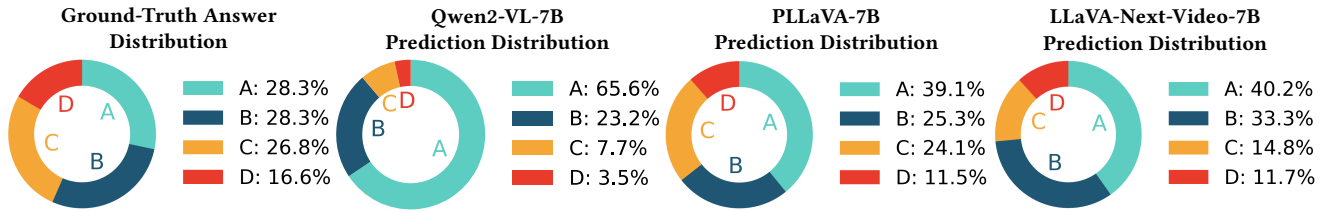


Figure 5: Distribution of Ground-Truth Answers and Sample VLLM Predictions: In DVBench, there exist some questions that offer only 2 or 3 answer choices, leading to a slightly uneven distribution of ground-truth answers.

from both evaluation strategies on DVBench. For most VLLMs, switching from IndividualEval to GroupEval results in a notable drop in accuracy. Generally, comparisons using GroupEval reveal a more substantial performance gap between different VLLMs. For example, Qwen2-VL-72B outperforms its 7B counterpart by 5.9% in Top-1 accuracy under IndividualEval, whereas the gap widens significantly to 18% under GroupEval. Nearly all VLLMs experience a performance drop of around 10% to 20% when moving from IndividualEval to the more demanding GroupEval, underscoring the latter’s challenge. For subsequent experiments, we adopt the rigorous and well-defined GroupEval as our default evaluation strategy.

Table 3: Top-1 accuracy of VLLMs on DVBench, evaluated with (w. DK) and without (w/o. DK) domain knowledge.

VLLMs	w/o. DK	w. DK	Δ
LLaMA-VID-7B [16]	6.4%	9.2%	+ 2.8%
LLaMA-VID-13B [16]	8.1%	8.8%	+ 0.7%
LLaVA-One-Vision-0.5B [14]	8.1%	12.7%	+ 4.6%
Qwen2-VL-7B [39]	10.6%	25.1%	+ 14.5%
LLaVA-Next-Video-7B [50]	15.5%	15.5%	+ 0.0%
Video-LLaVa-7B [17]	18.0%	18.4%	+ 0.4%
PLLaVA-7B [44]	17.3%	16.6%	- 0.7%
LLaVA-Next-Video-34B [50]	19.8%	18.0%	- 1.8%
LLaVA-One-Vision-7B [14]	21.2%	21.2%	+ 0.0%
PLLaVA-13B [44]	21.6%	16.7%	- 4.9%
Qwen2-VL-72B [39]	28.6%	32.9%	+ 4.2%
Qwen2-VL-2B [39]	26.9%	25.5%	- 1.4%
MiniCPM-V [47]	29.7%	35.7%	+ 6.0%
LLaVA-One-Vision-72B [14]	35.7%	36.8%	+ 1.1%
Average	19.1%	20.9%	+ 1.8%

4.4 Domain Knowledge Matters

In this subsection, we explore whether domain-specific knowledge can boost VLLM performance. Autonomous driving, in particular, involves concepts that may be unfamiliar to those outside the field, which can present a challenge for these models. To address this, we add clarifying information after each question to provide the necessary context. For instance, consider the question: “*What was the relation of the subject vehicle to the junction depicted in the video?*” While straightforward to experts, this phrasing may be unclear to individuals (or models) without driving experience. By providing domain knowledge, such as: “*Relation to Junction: The spatial (rather than causal) relationship of the subject vehicle to a junction at the*

beginning of the Precipitating Event. A junction is ... typically with differing travel speeds or directions.”, we make the question more accessible. This contextual information can help general-purpose VLLMs better understand the question, potentially improving their accuracy and effectiveness in such domain-specific tasks.

In Table 3, we report the top-1 accuracy of VLLMs on DVBench with and without domain knowledge. Most models show a significant accuracy boost with domain knowledge, highlighting the value of contextual understanding. However, certain models, such as PLLaVA-13B [44], experience a slight decline in domain knowledge. We believe this may stem from factors such as misinterpretation of domain-specific terminology or limitations in the model’s ability to generalize under the added complexity of domain-focused information. This suggests that while domain knowledge enhances understanding, its integration must be carefully calibrated to avoid adverse effects on certain models.

4.5 Main Results

Overall Evaluation. We evaluated various VLLMs on DVBench across two key dimensions: Perception and Reasoning, each with distinct sub-categories aligned with autonomous driving evaluation requirements. L2 ability evaluation results are shown in Table 4, while more fine-grained L3 results appear in Figure 1. The LLaVA-One-Vision-72B and MiniCPM-V achieved the highest overall scores, with perception accuracies of 39.4% and 36.7%, and reasoning accuracies of 35.4% and 36.4%, respectively. Notably, no model surpassed 40% overall accuracy, highlighting a significant gap between current VLLM capabilities and the stringent requirements for real-world autonomous driving. In such applications, even minor errors pose serious safety risks, underscoring the need for extensive domain-specific adaptation and training.

Perception Outperforms Reasoning. In the Perception tasks, Qwen2-VL-72B achieved the highest score in Environmental Conditions (EC) with 76.7%, followed by Video-LLaVA-7B and LLaVA-One-Vision-7B at 70.0%. For Physical Infrastructure (PI), only Qwen2-VL-72B and LLaVA-One-Vision-72B showed relatively strong performance. In Object (Obj) detection, MiniCPM-V led with 36.4%, closed followed by LLaVA-One-Vision-72B at 34.1%. For Reasoning tasks, LLaVA-One-Vision-72B scored highest in Risk & Hazard Assessment (RHA) with 46.6% and achieved 34.1% in Causal & Responsibility (CR), indicating some capacity for hazard assessment and causality attribution. However, overall reasoning performance remains limited. These results show that VLLMs excel in perception but struggle with reasoning—an essential capability for autonomous driving. The gap in reasoning highlights significant

Table 4: Detailed performance of the VLLMs on DVBench using GroupEval (L2 abilities). The following abbreviations are used: EC (Environmental Conditions), PI (Physical Infrastructure), OC (Operational Constraints), Obj (Objects), Zone (Zones); EU (Event Understanding), BMA (Behavior & Maneuver Analysis), SR (Spatial Reasoning), RHA (Risk & Hazard Assessment), CR (Causal & Responsibility). Additionally, the overall performance for each category (L1 abilities) is also reported.

VLLMs	Perception						Reasoning					
	Overall	EC	PI	OC	Obj	Zone	Overall	EU	BMA	SR	RHA	CR
LLaMA-VID-7B [16]	12.2%	26.7%	2.8%	13.5%	11.4%	9.1%	10.4%	2.0%	3.3%	7.5%	19.3%	11.4%
LLaMA-VID-13B [16]	12.8%	30.0%	0.0%	8.1%	15.9%	12.1%	9.6%	3.9%	0.0%	6.0%	21.6%	4.5%
LLaVA-One-Vision-0.5B [14]	15.0%	40.0%	2.8%	16.2%	13.6%	6.1%	11.1%	5.9%	10.0%	9.0%	21.6%	0.0%
Qwen2-VL-7B [39]	25.6%	43.3%	2.8%	27.0%	27.3%	30.3%	27.1%	33.3%	13.3%	17.9%	33.0%	31.8%
LLaVA-Next-Video-7B [50]	18.3%	43.3%	2.8%	16.2%	18.2%	15.2%	16.1%	7.8%	13.3%	10.4%	27.3%	13.6%
Video-LLaVa-7B [17]	21.1%	70.0%	2.8%	8.1%	22.7%	9.1%	18.6%	5.9%	13.3%	10.4%	34.1%	18.2%
PLLaVA-7B [44]	20.0%	56.7%	2.8%	10.8%	18.2%	18.2%	17.9%	3.9%	10.0%	10.4%	34.1%	18.2%
LLaVA-Next-Video-34B [50]	23.3%	50.0%	25.0%	21.6%	11.4%	15.2%	16.1%	5.9%	10.0%	17.9%	27.3%	6.8%
LLaVA-One-Vision-7B [14]	28.3%	70.0%	19.4%	21.6%	20.5%	18.2%	20.0%	3.9%	20.0%	16.4%	37.5%	9.1%
PLLaVA-13B [44]	23.9%	63.3%	11.1%	18.9%	18.2%	15.2%	15.4%	5.9%	0.0%	9.0%	34.1%	9.1%
Qwen2-VL-72B [39]	32.8%	50.0%	25.0%	35.1%	22.7%	36.4%	33.9%	41.2%	13.3%	31.3%	42.0%	27.3%
Qwen2-VL-2B [39]	31.7%	76.7%	27.8%	24.3%	20.5%	18.2%	26.4%	7.8%	26.7%	16.4%	44.3%	27.3%
MiniCPM-V [47]	39.4%	70.0%	19.4%	45.9%	36.4%	30.3%	35.4%	39.2%	16.7%	22.4%	53.4%	27.3%
LLaVA-One-Vision-72B [14]	36.7%	66.7%	16.7%	40.5%	34.1%	30.3%	36.4%	33.3%	30.0%	29.9%	46.6%	34.1%

challenges in real-world deployment, where robust, context-aware decision-making is critical for safety and reliability.

Foundation Models: A Key but Not Sole Factor in Success

Contrary to the common belief that larger models are always superior, DVBench results indicate that this is not necessarily the case. For instance, MiniCPM-V (8B parameters) outperforms the larger LLaVA-One-Vision-72B in perception tasks and performs comparably in reasoning tasks. We observed a trend that models using Qwen2 as the language backbone and SigLip [49] as the vision encoder tend to achieve strong performance. While this suggests that foundation models play a crucial role in success, they are not the sole determining factor, as model performance is influenced by a combination of architecture, training data, and optimization strategies. This highlights the need for a holistic approach to model design and training for autonomous driving applications.

4.6 Impact of Fine-Tuning on VLLMs

Thus far, our evaluation has focused on general-purpose VLLMs in safety-critical driving videos. While these models serve as foundation models that underpin downstream tasks, they are not optimized for autonomous driving. To assess their adaptability, we fine-tune VLLMs using safety-critical driving videos, measuring their post-fine-tuning performance and the improvements achieved.

Fine-Tuning Setup. To evaluate the impact of fine-tuning, we selected the Qwen2-VL series as our backbone model, as it covers a wide range of parameter scales and demonstrates strong performance among tested VLLMs. Specifically, we fine-tuned Qwen2-VL-2B and Qwen2-VL-7B on 2,880 question-answer pairs from DVBench, ensuring alignment with real-world driving scenarios.

Performance Gains from Fine-Tuning. Table 5 presents the accuracy comparison between fine-tuned and non-fine-tuned Qwen2-VL models. We observe notable improvements in accuracy, particularly for the Qwen2-VL-7B model, which exhibited a +10.94 percent points increase after fine-tuning, suggesting that larger models

benefit significantly from domain adaptation. Similarly, the Qwen2-VL-2B model showed a +5.24 percent points improvement, while Qwen2-VL-72B was excluded due to computational constraints.

Analysis and Implications. The results demonstrate that task-specific fine-tuning significantly enhances VLLMs’ comprehension of driving scenarios, reinforcing the necessity of domain-specific adaptation to bridge the gap between general-purpose visual reasoning and high-stakes decision-making in autonomous driving. While VLLMs exhibit strong baseline capabilities, our findings highlight that pre-trained models alone are not sufficiently reliable for safety-critical applications. Effective fine-tuning and optimization are essential to ensure deployment-ready accuracy, making adaptation strategies a crucial step toward improving VLLMs’ viability in real-world autonomous driving systems.

Table 5: Accuracy comparison of various VLLM models with and without fine-tuning.

VLLM Model	w/o. FT	w. FT	Δ
Qwen2-VL-2B	25.5%	30.74%	+5.24%
Qwen2-VL-7B	25.1%	36.04%	+10.94%

5 Conclusion and Future Work

We introduced DVBench, the first benchmark specifically designed to evaluate VLLMs in safety-critical driving video understanding. Built around a hierarchical ability taxonomy rooted in widely adopted frameworks for describing driving scenarios, DVBench structures its evaluation into three levels across 25 essential driving safety abilities. Our evaluation of 14 state-of-the-art VLLMs, ranging from 0.5B to 72B parameters, revealed substantial performance gaps, with no model surpassing 40% accuracy under our GroupEval strategy. Fine-tuning with domain-specific data achieves accuracy gains of up to 10.94 percentage points, underscoring the necessity

of targeted adaptation to bridge the gap between general-purpose VLLMs and mission-critical driving applications.

Future extensions of DVBBench will focus on broadening the evaluation landscape and increasing robustness. First, we plan to introduce additional question formats, such as True/False and open-ended questions, to better evaluate VLLMs' multi-step reasoning capabilities. Second, we aim to enrich DVBBench with more rare and long-tail driving scenarios—inspired by advances in rare pattern detection and long-tail learning across modalities and tasks [37, 38, 41, 51, 52]. Third, as VLLMs are increasingly integrated into safety-critical applications, we recognize the growing importance of trustworthiness, including robustness, reliability, and fairness [40, 42, 48]. Additionally, incorporating external knowledge and adopting multi-agent frameworks may further enhance model reasoning capabilities [33]. By continuously evolving DVBBench, we aim to establish a rigorous benchmark that catalyzes progress in VLLM-based understanding of complex and safety-critical driving environments.

Acknowledgements

We thank the anonymous reviewers for their constructive comments. This work is supported by the National Surface Transportation Safety Center for Excellence under Award No. 238717, National Science Foundation under Award No. IIS-2339989 and No. 2406439, DARPA under contract No. HR00112490370 and No. HR001124S0013, U.S. Department of Homeland Security under Grant Award No. 17STCIN00001-08-00, Amazon-Virginia Tech Initiative for Efficient and Robust Machine Learning, Amazon AWS, Google, Cisco, 4-VA, Commonwealth Cyber Initiative, National Surface Transportation Safety Center for Excellence, and Virginia Tech. The views and conclusions are those of the authors and should not be interpreted as representing the official policies of the funding agencies or the government.

References

- [1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. 2022. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems* 35 (2022), 23716–23736.
- [2] G Audi, AG Volkswagen, et al. [n. d.]. The PEGASUS Method.
- [3] Xinyun Chen, Ryan A Chi, Xuezhi Wang, and Denny Zhou. 2024. Premise order matters in reasoning with large language models. *arXiv preprint arXiv:2402.08939* (2024).
- [4] Thierry Deruyttere, Simon Vandenhende, Dusan Grujicic, Luc Van Gool, and Marie-Francine Moens. 2019. Talk2car: Taking control of your self-driving car. *arXiv preprint arXiv:1909.10838* (2019).
- [5] Chaoyou Fu, Yuhang Dai, Yondong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. 2024. Video-MME: The First-Ever Comprehensive Evaluation Benchmark of Multi-modal LLMs in Video Analysis. *arXiv preprint arXiv:2405.21075* (2024).
- [6] Xiaowei Gao, Pengxiang Li, Xinke Jiang, James Haworth, Jonathan Cardoso-Silva, and Ming Li. 2023. DriveScenify: Boosting Driving Scene Understanding with Advanced Vision-Language Models. <https://github.com/pixel99/DSify>
- [7] Xianda Guo, Ruijun Zhang, Yiqun Duan, Yuhang He, Chenming Zhang, Shuai Liu, and Long Chen. 2024. Drivemllm: A benchmark for spatial understanding with multimodal large language models in autonomous driving. *arXiv preprint arXiv:2411.13112* (2024).
- [8] Wencheng Han, Dongqian Guo, Cheng-Zhong Xu, and Jianbing Shen. 2024. Dme-driver: Integrating human decision logic and 3d scene perception in autonomous driving. *arXiv preprint arXiv:2401.03641* (2024).
- [9] Jonathan M Hankey, Miguel A Perez, and Julie A McClafferty. 2016. *Description of the SHRP 2 naturalistic database and the crash, near-crash, and baseline data sets*. Technical Report. Virginia Tech Transportation Institute.
- [10] 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* (2024).
- [11] Sandesh Jain, Surendrabikram Thapa, Kuan-Ting Chen, A Lynn Abbott, and Abhijit Sarkar. 2024. Semantic Understanding of Traffic Scenes with Large Vision Language Models. In *2024 IEEE Intelligent Vehicles Symposium (IV)*. IEEE.
- [12] Jean Kaddour, Joshua Harris, Maximilian Mozes, Herbie Bradley, Roberta Raileanu, and Robert McHardy. 2023. Challenges and applications of large language models. *arXiv preprint arXiv:2307.10169* (2023).
- [13] Jinkyu Kim, Anna Rohrbach, Trevor Darrell, John Canny, and Zeynep Akata. 2018. Textual explanations for self-driving vehicles. In *Proceedings of the European conference on computer vision (ECCV)*. 563–578.
- [14] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. 2024. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326* (2024).
- [15] Kunchang Li, Yali Wang, Yanan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, et al. 2024. Mvbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 22195–22206.
- [16] Yanwei Li, Chengyao Wang, and Jiaya Jia. 2025. Llama-vid: An image is worth 2 tokens in large language models. In *European Conference on Computer Vision*.
- [17] Bin Lin, Yang Ye, Bin Zhu, Jiayi Cui, Munan Ning, Peng Jin, and Li Yuan. 2023. Video-llava: Learning united visual representation by alignment before projection. *arXiv preprint arXiv:2311.10122* (2023).
- [18] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024. Visual instruction tuning. *Advances in neural information processing systems* 36 (2024).
- [19] Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics* 12 (2024), 157–173.
- [20] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. 2025. Mmbench: Is your multi-modal model an all-around player?. In *European Conference on Computer Vision*. Springer, 216–233.
- [21] Yuhang Lu, Yichen Yao, Jiadong Tu, Jianguan Shao, Yuexin Ma, and Xinge Zhu. 2024. Can LLMs Obtain a Driver's License? A Benchmark Towards Reliable AGI for Autonomous Driving. *arXiv preprint arXiv:2409.02914* (2024).
- [22] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. 2023. Video-chatgpt: Towards detailed video understanding via large vision and language models. *arXiv preprint arXiv:2306.05424* (2023).
- [23] Srikanth Malla, Chiho Choi, Isht Dwivedi, Joon Hee Choi, and Jiachen Li. 2023. Drama: Joint risk localization and captioning in driving. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*. 1043–1052.
- [24] Jiageng Mao, Yuxi Qian, Junjie Ye, Hang Zhao, and Yue Wang. 2023. Gpt-driver: Learning to drive with gpt. *arXiv preprint arXiv:2310.01415* (2023).
- [25] Ana-Maria Marcu, Long Chen, Jan Hünermann, Alice Karnsund, Benoit Hanotte, Prajwal Chidananda, Saurabh Nair, Vijay Badrinarayanan, Alex Kendall, Jamie Shotton, et al. 2024. LingoQA: Visual question answering for autonomous driving. In *European Conference on Computer Vision*. Springer, 252–269.
- [26] Ming Nie, Renyuan Peng, Chunwei Wang, Xinyue Cai, Jianhua Han, Hang Xu, and Li Zhang. 2025. Reason2drive: Towards interpretable and chain-based reasoning for autonomous driving. In *European Conference on Computer Vision*. Springer.
- [27] Tianwen Qian, Jingjing Chen, Linhai Zhuo, Yang Jiao, and Yu-Gang Jiang. 2024. Nuscene-qa: A multi-modal visual question answering benchmark for autonomous driving scenario. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 4542–4550.
- [28] Enna Sachdeva, Nakul Agarwal, Suhas Chundi, Sean Roelofs, Jiachen Li, Mykel Kochenderfer, Chiho Choi, and Behzad Dariush. 2024. Rank2tell: A multimodal driving dataset for joint importance ranking and reasoning. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*. 7513–7522.
- [29] Liang Shi, Boyu Jiang, Tong Zeng, and Feng Guo. 2025. ScVLM: Enhancing Vision-Language Model for Safety-Critical Event Understanding. In *Proceedings of the Winter Conference on Applications of Computer Vision*. 1061–1071.
- [30] Maged Shoman, Dongdong Wang, Armstrong Aboah, and Mohamed Abdel-Aty. 2024. Enhancing Traffic Safety with Parallel Dense Video Captioning for End-to-End Event Analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. 7125–7133.
- [31] Chonghao Sima, Katrin Renz, Kashyap Chitta, Li Chen, Hanxue Zhang, Chengen Xie, Jens Beifwenger, Ping Luo, Andreas Geiger, and Hongyang Li. 2024. Drivelm: Driving with graph visual question answering. In *European Conference on Computer Vision*. Springer, 256–274.
- [32] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, et al. 2020. Scalability in perception for autonomous driving: Waymo open dataset. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*.
- [33] Pawin Taechoyotin, Guanchao Wang, Tong Zeng, Bradley Sides, and Daniel Acuna. [n. d.]. MAMORX: Multi-agent multi-modal scientific review generation with external knowledge. In *NeurIPS 2024 Workshop Foundation Models for Science: Progress, Opportunities, and Challenges*.
- [34] 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* (2023).
- [35] Eric Thorn, Shawn C Kimmel, Michelle Chaka, Booz Allen Hamilton, et al. 2018. *A framework for automated driving system testable cases and scenarios*. Technical Report. United States. Department of Transportation. National Highway Traffic Safety
- [36] Xiaoyu Tian, Junru Gu, Bailin Li, Yicheng Liu, Yang Wang, Zhiyong Zhao, Kun Zhan, Peng Jia, Xianpeng Lang, and Hang Zhao. 2024. Drivevln: The convergence of autonomous driving and large vision-language models. *arXiv preprint arXiv:2402.12289* (2024).
- [37] Haohui Wang, Weijie Guan, Chen Jianpeng, Zi Wang, and Dawei Zhou. 2024. Towards heterogeneous long-tailed learning: Benchmarking, metrics, and toolbox. *Advances in Neural Information Processing Systems* 37 (2024), 73098–73123.
- [38] Haohui Wang, Baoyu Jing, Kaize Ding, Yada Zhu, Wei Cheng, Si Zhang, Yonghui Fan, Liqing Zhang, and Dawei Zhou. 2024. Mastering long-tail complexity on graphs: Characterization, learning, and generalization. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 3045–3056.
- [39] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. 2024. Qwen2-vl: Enhancing vision-language model’s perception of the world at any resolution. *arXiv preprint arXiv:2409.12191* (2024).
- [40] Longfeng Wu. 2024. Towards Trustworthy Graph Neural Networks and Their Applications in Recommender Systems. In *2024 IEEE International Conference on Big Data (BigData)*. IEEE, 8250–8252.
- [41] Longfeng Wu, Bowen Lei, Dongkuan Xu, and Dawei Zhou. 2023. Towards reliable rare category analysis on graphs via individual calibration. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*.
- [42] Longfeng Wu, Yao Zhou, Jian Kang, and Dawei Zhou. 2025. Bridging Fairness and Uncertainty: Theoretical Insights and Practical Strategies for Equalized Coverage in GNNs. In *Proceedings of the ACM on Web Conference 2025*. 4625–4634.
- [43] 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* (2025).
- [44] Lin Xu, Yilin Zhao, Daquan Zhou, Zhijie Lin, See Kiong Ng, and Jiashi Feng. 2024. Pllava: Parameter-free llava extension from images to videos for video dense captioning. *arXiv preprint arXiv:2404.16994* (2024).
- [45] Yiran Xu, Xiaoyin Yang, Lihang Gong, Hsuan-Chu Lin, Tz-Ying Wu, Yunsheng Li, and Nuno Vasconcelos. 2020. Explainable object-induced action decision for autonomous vehicles. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 9523–9532.
- [46] Zhenhua Xu, Yujia Zhang, Enze Xie, Zhen Zhao, Yong Guo, Kwan-Yee K Wong, Zhenguo Li, and Hengshuang Zhao. 2024. Drivegpt4: Interpretable end-to-end autonomous driving via large language model. *IEEE Robotics and Automation Letters* (2024).
- [47] 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* (2024).
- [48] Mang Ye, Xuankun Rong, Wenke Huang, Bo Du, Nenghai Yu, and Dacheng Tao. 2025. A survey of safety on large vision-language models: Attacks, defenses and evaluations. *arXiv preprint arXiv:2502.14881* (2025).
- [49] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 11975–11986.
- [50] Yuanhan Zhang, Bo Li, haotian Liu, Yong jae Lee, Liangke Gui, Di Fu, Jiashi Feng, Ziwei Liu, and Chunyuan Li. 2024. LLaVA-NeXT: A Strong Zero-shot Video Understanding Model.
- [51] Dawei Zhou, Jingrui He, K Selçuk Candan, and Hasan Davulcu. 2015. Muvir: Multi-view rare category detection. In *IJCAI*. Citeseer, 4098–4104.
- [52] Dawei Zhou, Jingrui He, Yu Cao, and Jae-sun Seo. 2016. Bi-level rare temporal pattern detection. In *2016 IEEE 16th International Conference on Data Mining (ICDM)*. IEEE, 719–728.
- [53] 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* (2023).
- [54] Yuchen Zhuang, Yue Yu, Kuan Wang, Haotian Sun, and Chao Zhang. 2023. Toolqa: A dataset for llm question answering with external tools. *Advances in Neural Information Processing Systems* 36 (2023), 50117–50143.

A Details of Hierarchical Ability Taxonomy

The DV Bench Ability Taxonomy consists of two fundamental components that capture the multifaceted nature of visual-linguistic understanding: perception and reasoning abilities. Perception abilities encompass foundational skills such as object recognition, environmental condition parsing, and infrastructure comprehension.

Reasoning abilities, on the other hand, involve sophisticated cognitive processes including inference, prediction, and causal analysis. By organizing the taxonomy along these two dimensions, DV Bench offers a comprehensive framework for evaluating how effectively Vision-Language Large Models interpret and synthesize complex visual-textual scenarios. The following subsections provide detailed definitions for each L3 ability within this taxonomy.

A.1 Perception

- i): Environmental Conditions (EC): Focuses on external factors affecting perception.
 - Atmospheric Conditions: Includes factors like Light and Weather that impact visibility.
 - Road Surface Condition: Essential for traction and vehicle control, could be classified as "Snowy", "Dry" or "Wet".
- ii): Physical Infrastructure (PI): Pertains to the static aspects of the road environment.
 - Road Geometry: Includes road alignment and grade affect vehicle dynamics.
 - Lane Configuration: Includes the number of contiguous travel lanes and the number of through travel lanes related to road layout.
- iii): Operational Constraints (OC): Dynamic elements constraining vehicle operation.
 - Traffic Conditions: Includes the factors such as the traffic flow and traffic density that influence driving decisions.
 - Traffic Control Devices: The traffic control includes signs and signals that dictate rules.
 - Visibility Constraints: Visual obstructions impact the driver’s line of sight.
- iv): Objects (Obj): Identifies entities the vehicle must detect and interpret.
 - Static Objects: Includes factors such as the number of objects or animals involved in an event, or the specification of other vehicles, pedestrians, animals, or objects that are involved in the event.
 - Other Road Users: How many other motorists or non-motorists were involved in the event depicted in the video?
- v): Zones (Zone): Specific areas with unique driving rules or conditions.
 - Geographic Zones: The description of surroundings that may influence the flow of traffic, such as Urban, School, Open Residential, Business, or other settings.
 - Intersection and Junctions: Includes factors such as the relation to the junction and the influence of the intersection on the vehicle’s movement.

A.2 Reasoning

- i): Event Understanding (EU): Comprehends the nature and context of events.
 - Event Characteristics: Comprehends factors from video, such as the type of incident or the nature of an event.
 - Event Triggers: The precipitating event that initiated the sequence leading to the crash or near-crash, i.e., identifies causes of the event.
 - Event Severity: Understand the severity of an event, if it is a crash, infer the severity of the crash.

Table 6: The vision encoder and language model details of the evaluated VLLMs.

VLLMs	Vision Encoder	Vision Encoder Size	Language Model	LLM Size
LLaMA-VID-7B [16]	EVA-G	1.1B	Vicuna	7B
LLaMA-VID-13B [16]	EVA-G	1.1B	Vicuna	13B
LLaVA-Next-Video-7B [50]	SigLIP	400M	Vicuna-v1.5	7B
LLaVA-Next-Video-34B [50]	SigLIP	400M	Vicuna-v1.5	34B
LLaVA-One-Vision-5B [14]	SigLIP	400M	Qwen-2	5B
LLaVA-One-Vision-7B [14]	SigLIP	400M	Qwen-2	7B
LLaVA-One-Vision-72B [14]	SigLIP	400M	Qwen-2	72B
MiniCPM-V [47]	SigLIP	400M	Qwen2	7B
PLLaVA-7B [44]	CLIP ViT-L	428M	Vicuna-v1.5	7B
PLLaVA-13 [44]	CLIP ViT-L	428M	Vicuna-v1.5	13B
Qwen2-VL-2B [39]	RoPE-2D+ViT	675M	Qwen2	1.5B
Qwen2-VL-7B [39]	RoPE-2D+ViT	675M	Qwen2	7.6B
Qwen2-VL-72 [39]	RoPE-2D+ViT	675M	Qwen2	72B
Video-LLaVa-7B [17]	OpenCLIP-L/14	428M	Vicuna-v1.5	7B

Table 7: Detailed performance of the VLLMs on DVBench using GroupEval without domain knowledge(L2 abilities). The following abbreviations are used: EC for Environmental Conditions; PI for Physical Infrastructure; OC for Operational Constraints; Obj for Objects; and Zone for Zones; EU for Event Understanding; BMA for Behavior & Maneuver Analysis; SR for Spatial Reasoning; RHA for Risk & Hazard Assessment; CR for Causal & Responsibility. Additionally, the results of L1 abilities are also reported, namely, the overall performance under each category correspondingly.

VLLMs	Perception						Reasoning					
	Overall	EC	PI	OC	Obj	Zone	Overall	EU	BMA	SR	RHA	CR
LLaMA-VID-7B [16]	7.8%	20.0%	5.6%	5.4%	0%	12.1%	6.6%	7.8%	0%	6.5%	10.2%	2.3%
LLaMA-VID-13B [16]	11.1%	20.0%	11.1%	5.4%	9.1%	12.1%	8.6%	5.9%	0%	10.4%	13.6%	4.5%
LLaVA-One-Vision-5B [14]	11.7%	36.7%	0%	5.4%	11.4%	9.1%	7.2%	0%	6.7%	2.6%	19.3%	0%
Qwen2-VL-7B [39]	13.3%	20.0%	13.9%	5.4%	13.6%	15.2%	11.7%	3.9%	3.3%	11.7%	17.0%	15.9%
LLaVA-Next-Video-7B [50]	16.1%	40.0%	2.8%	13.5%	11.4%	18.2%	15.2%	9.8%	16.7%	10.4%	22.7%	13.6%
Video-LLaVa-7B [17]	18.9%	63.3%	5.6%	13.5%	11.4%	9.1%	20.7%	11.8%	20.0%	7.8%	31.8%	31.8%
PLLaVA-7B [44]	20.6%	50.0%	2.8%	13.5%	20.5%	21.2%	18.6%	11.8%	10.0%	11.7%	33.0%	15.9%
LLaVA-Next-Video-34B [50]	23.9%	53.3%	22.2%	24.3%	11.4%	15.2%	19.0%	9.8%	13.3%	18.2%	29.5%	13.6%
LLaVA-One-Vision-7B [14]	26.7%	70.0%	19.4%	21.6%	13.6%	18.2%	20.3%	9.8%	16.7%	16.9%	36.4%	9.1%
PLLaVA-13B [44]	27.2%	60.0%	16.7%	21.6%	27.3%	15.2%	21.0%	11.8%	13.3%	16.9%	36.4%	13.6%
Qwen2-VL-72B [39]	28.3%	53.3%	22.2%	32.4%	20.5%	18.2%	30.3%	33.3%	13.3%	24.7%	42.0%	25.0%
Qwen2-VL-2B [39]	33.3%	83.3%	30.6%	24.3%	22.7%	15.2%	27.2%	17.6%	10.0%	20.8%	47.7%	20.5%
MiniCPM-V [47]	35.0%	73.3%	13.9%	32.4%	31.8%	30.3%	28.6%	29.4%	10.0%	18.2%	48.9%	18.2%
LLaVA-One-Vision-72B [14]	38.3%	70.0%	19.4%	40.5%	34.1%	33.3%	34.5%	27.5%	26.7%	31.2%	47.7%	27.3%

- ii): Behavior & Maneuver Analysis (BMA): Evaluates actions taken by the vehicle.
 - Vehicle Maneuvers: Analyze the vehicle responses before, on, and after the safety-critical event, such as the pre-incident maneuver, the evasive maneuver, and the post-maneuver control.
 - Maneuver Evaluation: The maneuver judgment assesses decision quality.
- iii): Spatial Reasoning (SR): Understands spatial relationships and positioning.
 - Road and Terrain Features: Infer the road alignment, grade, and locality that impact navigation.
 - Lane Positioning: Comprehends the context related to lane usage, such as which lane was the subject vehicle in during the event in the video.
- Spatial Relationships: Reason about the spatial relationships that help in situational awareness, for example, where was the other vehicle, pedestrian, animal, or object located in relation to the subject vehicle.
- iv): Risk & Hazard Assessment (RHA): Identifies and evaluates potential dangers.
 - Environmental Hazards: Understanding the environmental factors that affect safety.
 - Traffic Hazards: Reason on the factors that might pose risks to the subject vehicle.
 - Obstruction Hazards: Infer from the video what obstructions may require evasive actions.
- v): Causal & Responsibility (CR): Determines cause and responsibility of a safety-critical event.

MCQ Prompt for Zero-shot Inference

Question: {question_text}

Options:
 A. xxx
 B. xxx
 C. xxx [optional]
 D. xxx [optional]

Domain Knowledge: {domain_knowledge}

Please select the best option that accurately addresses the question and returns only the option letter.

Figure 7: The prompt template for zero-shot VLLM inference.

- Fault Analysis: Assesses liability, i.e., which driver or non-motorist (if any) committed an error that led to the event.
- Cause and Effect: Explain why an event occurred.

B Model Configurations

In Table 6, we present a comprehensive overview of all VLLMs evaluated in DVBBench. This table includes specific details on each model’s language backbone, vision encoder, and the corresponding number of parameters, offering insight into the architecture and scale of each model evaluated.

C Model Performance without Domain Knowledge

In this section, we give detailed results about the performance of various VLLMs on DVBBench without domain knowledge. Specifically, We present the evaluation results of L1 and L2 abilities in Table 7. Comparing these results with those obtained using domain knowledge, we observe a significant improvement in accuracy when domain knowledge is integrated. This observation aligns with the findings discussed in Section 4.4.

System Prompt

As an AI assistant for video analysis, your role is to examine footage from the front dash camera of an autonomous vehicle and accurately answer multiple-choice questions based on the video.

You will be provided with {num_frames} separate frames uniformly sampled from the video, the frames are provided in chronological order of the video. Please analyze these images and provide the answer to the following question about the video content.

Instructions:

1. Carefully analyze the objects, environment, and events within the video to understand its context and details.
2. Select the correct answer from the provided options by identifying only the option letter (e.g., A, B, C, or D).
3. Respond solely with the correct option letter, ensuring precision in your choice based on the observed content.

Based on your observations, select the best option that accurately addresses the question.

Figure 6: The system prompt for VLLM answer generation in DVBBench.

D Prompt Templates

To facilitate a fair comparison of model performance, all experiments in this paper are conducted using zero-shot inference. Figure 6 and Figure 7 illustrate the System Prompt and MCQ Question-Answering Prompt employed in the performance evaluation.

E Detailed Statistics of the MCQ Bank

To further analyze the coverage and distribution of our Multiple-Choice Question (MCQ) bank, we provide a breakdown of the MCQ counts for each ability. The statistics are structured hierarchically, encompassing Level 1 categories (Perception and Reasoning), Level 2 subcategories (e.g., Environmental Conditions, Objects, and Zones), and Level 3 specific abilities (e.g., Atmospheric Conditions, Vehicle Maneuvers, and Traffic Hazards).

As shown in Table 8, the MCQ bank spans 25 distinct Level 3 abilities with a total of approximately 10,000 questions. Each level refines the scope of assessment, enabling targeted evaluation of specific capabilities. These statistics offer a clear view of the MCQ bank’s depth and its potential to systematically benchmark VLLMs’ understanding of safety-critical and normal driving scenarios.

Table 8: The Number of Multiple-Choice Questions (MCQs) for Each Ability Across Different Levels

Ability	Level	# MCQs
Perception	L1	3925
Reasoning	L1	6338
Perception/Environmental Conditions	L2	603
Perception/Objects	L2	999
Perception/Operational Constraints	L2	721
Perception/Physical Infrastructure	L2	802
Perception/Zones	L2	800
Reasoning/Behavior & Maneuver Analysis	L2	802
Reasoning/Causal and Responsibility	L2	999
Reasoning/Event Understanding	L2	1008
Reasoning/Risk & Hazard Assessment	L2	1924
Reasoning/Spatial Reasoning	L2	1605
Perception/Environmental Conditions/Atmospheric Conditions	L3	402
Perception/Environmental Conditions/Road Surface Conditions	L3	201
Perception/Objects/Other Road Users	L3	602
Perception/Objects/Static Objects	L3	397
Perception/Operational Constraints/Traffic Conditions	L3	401
Perception/Operational Constraints/Traffic Control Devices	L3	198
Perception/Operational Constraints/Visibility Constraints	L3	122
Perception/Physical Infrastructure/Lane Configuration	L3	400
Perception/Physical Infrastructure/Road Geometry	L3	402
Perception/Zones/Construction Zones	L3	199
Perception/Zones/Geographic Zones	L3	200
Perception/Zones/Intersection and Junctions	L3	401
Reasoning/Behavior & Maneuver Analysis/Maneuver Evaluation	L3	200
Reasoning/Behavior & Maneuver Analysis/Vehicle Maneuvers	L3	602
Reasoning/Causal and Responsibility/Cause and Effect	L3	598
Reasoning/Causal and Responsibility/Fault Analysis	L3	401
Reasoning/Event Understanding/Event Characteristics	L3	411
Reasoning/Event Understanding/Event Severity	L3	401
Reasoning/Event Understanding/Event Triggers	L3	196
Reasoning/Risk & Hazard Assessment/Environmental Hazards	L3	725
Reasoning/Risk & Hazard Assessment/Obstruction Hazards	L3	600
Reasoning/Risk & Hazard Assessment/Traffic Hazards	L3	599
Reasoning/Spatial Reasoning/Lane Positioning	L3	401
Reasoning/Spatial Reasoning/Road and Terrain Features	L3	602
Reasoning/Spatial Reasoning/Spatial Relationships	L3	602