# Can LVLMs Obtain a Driver's License? A Benchmark Towards Reliable AGI for Autonomous Driving

**Yuhang Lu**[1,*], **Yichen Yao**[1,*], **Jiadong Tu**[1,*], **Jiangnan Shao**[1,*], **Yuexin Ma**[1,†], **Xinge Zhu**[2,†]
[1]ShanghaiTech University
[2]Shanghai Jiao Tong University
{luyh2, yaoych2023, tujd2023, shaojn2023, mayuexin}@shanghaitech.edu.cn, zhuxinge123@gmail.com

## Abstract

Large Vision-Language Models (LVLMs) have recently garnered significant attention, with many efforts aimed at harnessing their general knowledge to enhance the interpretability and robustness of autonomous driving models. However, LVLMs typically rely on large, general-purpose datasets and lack the specialized expertise required for professional and safe driving. Existing vision-language driving datasets focus primarily on scene understanding and decision-making, without providing explicit guidance on traffic rules and driving skills, which are critical aspects directly related to driving safety. To bridge this gap, we propose IDKB, a large-scale dataset containing over one million data items collected from various countries, including driving handbooks, theory test data, and simulated road test data. Much like the process of obtaining a driver's license, IDKB encompasses nearly all the explicit knowledge needed for driving from theory to practice. In particular, we conducted comprehensive tests on 15 LVLMs using IDKB to assess their reliability in the context of autonomous driving and provided extensive analysis. We also fine-tuned popular models, achieving notable performance improvements, which further validate the significance of our dataset.

**Project Page** — 4dvlab.github.io/project_page/idkb.html

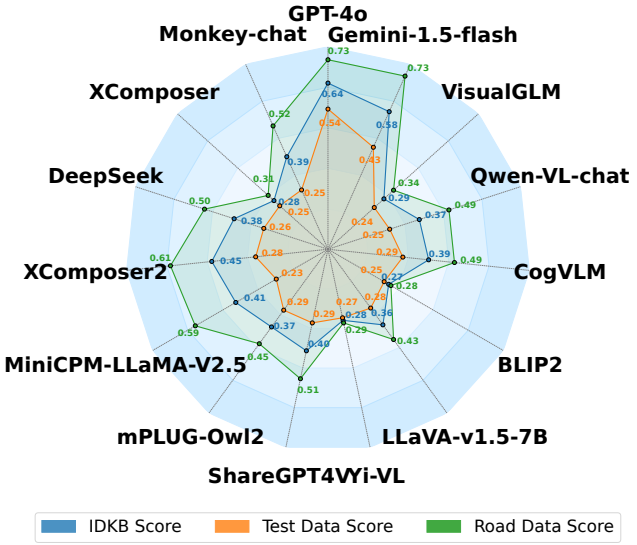



Figure 1: Performance of 15 representative Large Vision-Language Models on IDKB, evaluated by three driving knowledge understanding metrics.

## 1 Introduction

In recent years, Large Vision-Language Models (LVLMs) (Achiam et al. 2023; Bai et al. 2023; Liu et al. 2024) have emerged as powerful tools in AI, showcasing impressive capabilities in areas such as visual dialogue and document understanding. Building on the general knowledge of LVLMs, some approaches (Xu et al. 2023; Mao et al. 2023; Sima et al. 2023) have leveraged these models to enhance the efficiency, robustness, and interpretability of autonomous vehicles, addressing the intricate challenges of autonomous driving in open world. However, LVLMs are often trained on vast and generic datasets, lacking the specialized expertise required for the driving domain. This gap in domain-specific knowledge can lead to potential inaccuracies when these models are applied to self-driving systems, where precision and reliability are paramount.

To address this issue, many vision-language driving datasets (Qian et al. 2024; Sima et al. 2023; Park et al. 2024; Kim et al. 2018; Malla et al. 2023; Tian et al. 2024; Li et al. 2024) towards LVLM fine-tuning have been developed. Most of these datasets simply add textual annotations to traffic images from existing datasets, which limits the complexity and diversity of the scenarios they cover. While a few datasets (Tian et al. 2024; Li et al. 2024) are specifically collected and annotated with more challenging driving scenarios, they still primarily focus on tasks like scene perception and decision-making, rather than providing structured driving knowledge. As a result, models built on these datasets can only implicitly learn driving knowledge through the supervision of driving decisions. However, this approach differs significantly from how humans learn to drive, which involves studying driving instructions, traffic laws, driving rules, driving skills, and methods for handling emergency

[*]These authors contributed equally.
[†]Corresponding authors.

situations. Consequently, these models often lack a comprehensive understanding of driving knowledge, leading to unstable and unreliable performance in real-world applications.

To this end, we propose **I**ntelligent **D**riving **K**nowledge **B**ase (IDKB), the first large-scale vision-language dataset dedicated to professional driving knowledge and experience. Typically, humans learn to drive systematically by studying driving materials, taking theory tests, and practicing on the road. To enable LVLMs to effectively "earn a driver's license" and guarantee their driving safety, we compiled an extensive collection of driving handbooks and test questions from various countries, covering traffic laws, rules, driving techniques, and crisis management skills. In addition to theoretical data, we generated practical data by simulating diverse road scenarios in CARLA (Dosovitskiy et al. 2017), including variations in weather, lighting, traffic conditions, and more. This comprehensive effort resulted in a dataset of over 1 million data entries with various formats, spanning 15 countries, 9 languages, and 4 vehicle types. By capturing diverse driving regulations and practices from various regions, our dataset provides a solid foundation for thoroughly evaluating LVLMs and enhancing their ability to acquire safe and efficient driving capabilities.

Based on IDKB, we conducted an extensive evaluation for 15 existing LVLMs to assess their degree of mastery in driving knowledge and skills. As shown in Fig. 1, the evaluated LVLMs generally lacked strong driving domain knowledge, underscoring the need for fine-tuning with high-quality, structured, and diverse driving knowledge data for effective application in autonomous driving.

We also fine-tuned several of these models using our dataset, and the experimental results demonstrate that explicit and structured driving knowledge significantly enhances the performance of LVLMs, leading to more effective and accurate outcomes. Our findings highlight the importance of incorporating specialized, domain-specific knowledge into LVLMs to better equip them for the complex and safety-critical task of autonomous driving.

Our key contributions can be summarized as follows:

- We introduce IDKB, the first large-scale vision-language dataset explicitly containing both driving theory and practical knowledge.
- We evaluate 15 existing LVLMs on our dataset and provide a comprehensive analysis of their driving abilities.
- We offer fine-tuned, open-source LVLMs trained on our dataset, which possess enhanced professional driving expertise.

## 2 Related Work

### LVLMs for Autonomous Driving

Recently, research on Large Vision-Language Models (LVLMs) has surged, with multimodal models such as GPT-4V (Achiam et al. 2023), Qwen (Bai et al. 2023), and LLaVA (Liu et al. 2024) demonstrating strong performance across a wide range of general tasks. Leveraging these generalized capabilities, several approaches have begun to integrate LVLMs with autonomous driving algorithms to enhance self-driving car performance and interpretability. For example, DriveGPT4 (Xu et al. 2023) processes multimodal input data and generates both text responses and vehicle control signals by fine-tuning a LVLM on an instruction-tuning dataset. AgentDriver (Mao et al. 2023) converts driving situations into textual descriptions with human-like intelligence, then uses an LLM to reason and plan. Similarly, DriveVLM (Sima et al. 2023) employs a LVLM to output planning trajectories through a Chain-of-Thought (CoT) reasoning process. However, due to being trained on vast amounts of general data, LVLMs often lack the specialized driving knowledge necessary for accuracy and reliability in driving. Therefore, a dataset that covers specific and comprehensive driving knowledge is crucial for both evaluating and enhancing LVLMs for autonomous driving.

### Vision-Language Driving Datasets

With the rise of Large Vision-Language Models (LVLMs), numerous vision-language datasets for autonomous driving have been developed for better understanding driving scenes. The pioneering works BDD-X (Kim et al. 2018) and BDD-OIA (Xu et al. 2020) annotated video datasets with textual descriptions and explanations of ego car actions. Many subsequent multimodal datasets have relabeled existing self-driving datasets. Talk2Car (Deruyttere et al. 2019) adds free-form, high-quality natural language commands to the nuScenes dataset. NuScenes-QA (Qian et al. 2024) creates 460,000 question-answer pairs based on 3D object relationships to evaluate models' understanding and reasoning abilities. DriveLM (Sima et al. 2023) constructs perception, prediction, and planning question-answer pairs in a graph structure to simulate human reasoning, thus enhancing end-to-end autonomous driving systems. VLAAD (Park et al. 2024) utilizes GPT-4 to generate Q&A pairs from BDD-X (Kim et al. 2018), producing an instruction-following dataset that features complex reasoning, detailed descriptions, and conversation. However, these datasets rely heavily on existing datasets and often lack complex scenarios. To address this, some datasets are collected and annotated from scratch. DRAMA (Malla et al. 2023) identifies critical objects in traffic scenarios and provides corresponding linguistic descriptions of driving risks. DriveVLM (Tian et al. 2024) presents SUP-AD, a scene understanding and planning dataset with annotations on challenging and long-tail scenarios. CODA-LM (Li et al. 2024) is a large-scale multimodal self-driving dataset focusing on road corner cases.

However, these works lack explicit knowledge of traffic regulations, rules, and driving techniques, limiting LVLMs in their ability to develop a comprehensive and abstract understanding of driving knowledge. In contrast, our dataset mirrors the human process of acquiring driving knowledge by collecting detailed annotations from driving handbooks and test questions, while also integrating theoretical learning with practical application through simulated road scenarios in CARLA. By explicitly presenting this knowledge, our approach closely aligns with human learning styles, enabling LVLMs to acquire and integrate driving knowledge more efficiently and reliably, ultimately enhancing their per-

formance in driving-related tasks.

## 3 IDKB Dataset

**I**ntelligent **D**riving **K**nowledge **B**ase (IDKB) is structured as a driving knowledge resource, mirroring the process individuals follow to acquire expertise when obtaining a driver's license. This process typically involves studying driving handbooks, taking theory tests, and practicing on the road. In this section, we introduce the data construction pipeline3 and present the statistics and characteristics of our dataset.

### Data Construction

**Driving Handbook Data** Driving handbooks are highly structured and comprehensive resources, covering laws, regulations, techniques, safety, and more. They serve as the foundational step in learning to drive. By studying these handbooks, an intelligent system can develop a thorough and well-rounded understanding of the driving domain. We collected 206 documents, including traffic laws and driving handbooks, totalling 23,847 pages from 15 different countries via the Internet. As shown in Fig. 2, we collect and organize unordered data from these documents in systematic way. We first employ layout detector and Optical Character Recognition (OCR) technology to extract data blocks and the text within these blocks. Subsequently, we developed an algorithm to cluster and sequence the data blocks in a way that aligns with human readability. Finally, we filter out duplicate and irrelevant data. An example is presented in Fig. 3.

**Driving Test Data** Driving test data represent an alternative format for the knowledge covered in driving handbooks. While the handbooks offer a structured and comprehensive overview, the test questions reorganize this knowledge into multiple-choice and short-answer formats. This approach allows an intelligent system to both reinforce what it has learned and assess its understanding in a more interactive manner. Engaging with these questions ensures that the system has thoroughly internalized the handbook content, making this process an essential part of learning. To construct this part of data, we extensively collected questions from driving tests of 15 different countries. As depicted in Fig. 2, the data collection process is similar with that of driving handbook data. We extract relevant information from various driving tests, and then reorganize these metadata into standard question-answer formats. Most entries we collected are multiple-choice questions, with one or more correct answers, while a smaller portion belongs to open QA questions. Two annotated examples are presented in Fig. 3.

**Data Augmentation.** To enhance the diversity of data and expand the scale of the dataset, we employed GPT-4o model to augment Driving Test Data. To ensure the enhanced data is well-structured, we divide the item into the question stem, options, and explanation sections. Each section undergoes incremental enhancement three times using GPT-4o. After completing the enhancements, the sections are combined to form the enhanced question-answer pair. To assure quality and avoid duplication, we require GPT-4o to output three distinct enhanced versions of the data in a single response, following a specific format. Invalid data is identified and removed through format checking and manual screening afterwards. In addition, for each data entry containing images, we generate textual descriptions of the images using GPT-4o and incorporate these descriptions into the dataset for further application.

**Data Quality Control.** To ensure high data quality, we implemented a two-step verification process at the end of the collection pipeline. Initially, an automated program filters out obvious low-quality data, such as images with extremely low resolution or very short text. After this automated removal, we conduct a manual review to further refine and ensure the quality of the remaining data.

**Driving Road Data** After studying the driving handbooks and test questions, the intelligent system gains sufficient theoretical knowledge of driving. The next crucial step is to apply and reinforce this knowledge in real-world driving scenarios, which ensures that the system can effectively translate theoretical understanding into practical, real-world competence. However, existing datasets for real-world driving scenarios are often limited in scope, scale, and coverage of traffic signs, making it difficult to thoroughly understand road conditions. To address these limitations, we leverage the CARLA (Dosovitskiy et al. 2017) simulator to generate a large, high-quality dataset that offers a more comprehensive understanding of traffic regulations at a low cost. Our approach begins with constructing custom simulated environments in the CARLA simulator to generate traffic sign understanding data. We further expand this dataset by extracting scenes with traffic signs from the Bench2Drive (Jia et al. 2024) dataset, thereby creating additional annotated traffic sign data. As demonstrated in Fig. 2, our driving road data collection process involves the following three steps.

**Scene Construction.** In the scene construction stage, we first collect high-resolution traffic signs from different countries and create 3D models for each traffic sign using the CARLA-UE4 editor. Then we select two of CARLA's large maps and set the traffic signs at appropriate locations within the CARLA simulator.

**Camera Data Collection.** In the camera data collection stage, we generate an ego vehicle equipped with camera sensors to collect image data by driving the vehicle manually in the CARLA simulator. To ensure the authenticity of the simulation data, we randomly generate a large number of vehicles and pedestrians in the CARLA world and control them through the Autopilot mode and the WalkerAIController provided by the CARLA, respectively, to simulate the real road conditions. To ensure the diversity and richness of the data, we set different weather conditions, times of day, and road surface slipperiness to simulate real camera views and driving scenes. While collecting the camera sensor data, we also record the position, bounding box, rotation and other necessary information of each actor in the CARLA world for the next step. In total, we drive the ego vehicle in the CARLA simulator for about 20 hours and obtain approximately 400,000 frames of camera data.

**Data Annotation.** After obtaining the image data and actor information, we automatically pick out the frames containing traffic signs based on the position, distance and di-

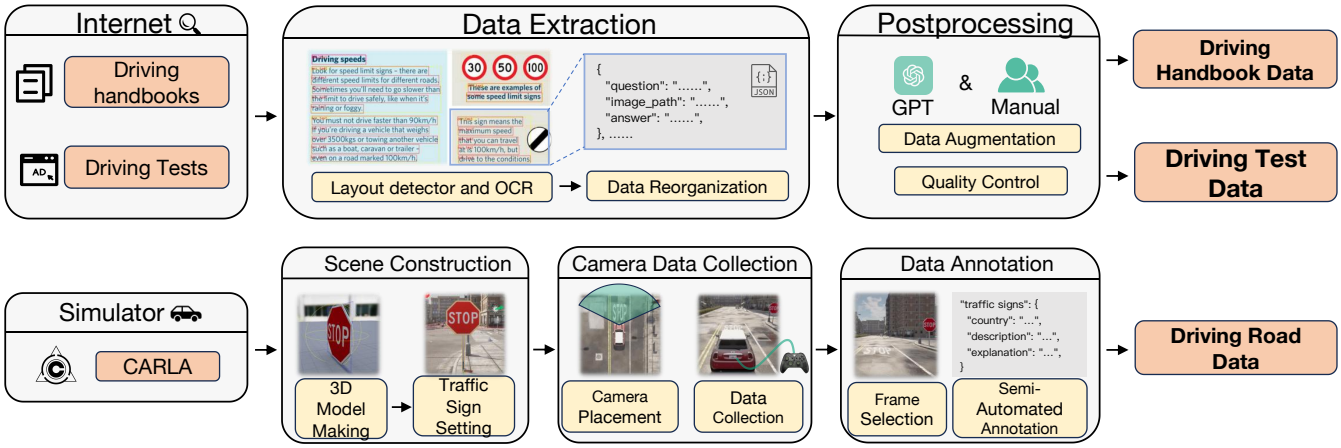



Figure 2: Data construction pipeline of IDKB dataset. For Driving Handbook and Driving Test Data, we collect comprehensive driving knowledge resources from internet, followed by data extraction and postprocessing to obtain the final data. For Driving Road Data, we utilize CARLA to generate simulated road scenarios focused on traffic sign comprehension.

| Dataset | Data Type | | Data Source | | Data Domain | | | Knowledge Domain | | | | Amount |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | QA | MCQ | Real | Syn. | Country | Language | Veh. Type | Laws | Signs | Techniques | Def. Drive | |
| BDD-X (Kim et al. 2018) | ✗ | ✗ | ✓ | ✗ | US | EN | Car | ✗ | ✓ | ✓ | ✗ | 26K |
| Talk2Car (Deruyttere et al. 2019) | ✗ | ✗ | ✓ | ✗ | US, SG | EN | Car | ✗ | ✗ | ✗ | ✗ | 12K |
| nuScenes-QA (Qian et al. 2024) | ✓ | ✗ | ✓ | ✗ | US, SG | EN | Car | ✗ | ✗ | ✗ | ✗ | 460K |
| DriveLM (Sima et al. 2023) | ✓ | ✗ | ✓ | ✓ | US, SG | EN | Car | ✗ | ✓ | ✓ | ✗ | 2M |
| DRAMA (Malla et al. 2023) | ✗ | ✗ | ✓ | ✗ | JP | EN | Car | ✗ | ✓ | ✓ | ✗ | 102K |
| LangAuto (Shao et al. 2024) | ✗ | ✗ | ✗ | ✓ | US | EN | Car | ✗ | ✓ | ✓ | ✗ | 64K |
| SUP-AD (Tian et al. 2024) | ✗ | ✗ | ✓ | ✗ | CN | EN | Car | ✗ | ✓ | ✓ | ✗ | - |
| VLAAD (Park et al. 2024) | ✓ | ✗ | ✓ | ✗ | US | EN | Car | ✗ | ✓ | ✓ | ✗ | 64K |
| CODA-LM (Li et al. 2022) | ✓ | ✗ | ✓ | ✗ | DE,CN,SG | EN | Car | ✗ | ✓ | ✓ | ✗ | 10K |
| **IDKB** | ✓ | ✓ | ✓ | ✓ | 15 | 9 | 4 | ✓ | ✓ | ✓ | ✓ | 1M |

Table 1: Comparison between our dataset and existing vision-language autonomous driving datasets. “QA” means “Question and Answer” and “MCQ” indicates “Multiple Choice Question”.

rection of the sign relative to the ego vehicle. Then we manually build a dictionary to define the descriptions and explanations for each sign, and attach the text annotations to each frame in an automated process. Finally, we obtained a total of 112,388 data samples, including both multiple-choice and question-and-answer formats.

**Data Quality Control.** Images that lack traffic signs or are obscured by adverse weather conditions or occlusions are considered as low-quality data. To maintain high data quality, we manually check and remove any images that do not meet the required standards.

## Dataset Statistics

In total, IDKB provides 1,016,956 data entries. As shown in Fig. 4(a), Driving Test Data constitutes the largest portion of the dataset (84.0%), with CARLA and Driving Handbook Data accounting for 11.1% and 5.0%, respectively. Our dataset spans multiple countries and multiple languages. As presented in Fig. 4(b)(c), apart from English-speaking countries, we also include driving knowledge from China, Italy, Germany, and others. In terms of vehicle types, we categorized the data into four classes: Car (standard passenger vehicles including sedan, jeep, ...), Truck (large vehicles including minivan, commercial, LGV, ...), Bus (including minibus, trailer, coaches, ...), and Moto (Motobike, Motocycle). Fig. 4(d) presents the distribution of the vehicle types.

To better analyze the knowledge coverage of our dataset, we employed proprietary LVLMs to classify all the questions into four major categories according to their semantics, including Laws & Regulations (22.2%), Road Signs & Signals (38.6%), Driving Techniques (22.0%), and Defensive Driving (17.1%). More data details are provided in Supplementary.

## Data Characteristics

As shown in Tab. 1, compared with existing vision-language autonomous driving datasets, IDKB possesses four main novel characteristics as follows.

**Diverse Data Type** Our dataset contains diverse data types, encompassing both Question and Answer (QA) and Multiple-choice Question (MCQ). Most existing datasets

## Driving Handbook Data

**Question:** "I will provide a subtitle from the handbook of *Australia* for *car, moto and truck* (maybe also give you some related images). Your task is to generate detailed content related to this subtitle. Ensure that the response includes specific information and guidance to help learners better understand and apply the knowledge.The subtitle is: *Pedestrian crossings*. "

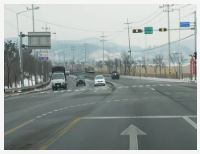
**Answer:** "Pedestrian crossings are marked by white stripes on the roadway and special signs. Motorists must give way to pedestrians and cyclists on a marked pedestrian crossing. Note: …"

**Country:** Australia
**Language:** English
**Vehicle:** Car/Truck/ Bus/Moto


## Driving Test Data

**Question:** "What are the two safest ways to drive in this situation?"
**Options:** A: Drive slowly while paying attention to the movement of surrounding vehicles.
B: You pass straight through the intersection.
C: Check if there are any pedestrians trying to cross the crosswalk.
D: Vehicles making a left turn ahead proceed without caution.
**Answer:** Option A, Option C

**Country:** Korea
**Language:** English
**Vehicle:** Car



---

**Question:** "¿Qué debe hacer si se encuentra con un semáforo de franja vertical blanca no-intermitente mientras conduce su autobús de largo recorrido por un carril reservado para autobuses? "
**Answer:** "Obedecer sus indicaciones. "

**Country:** Spain
**Language:** Spainish
**Vehicle:** Bus


## Driving Road Data

**Question:** "Assume you are driving a *car* in *America*. What is the traffic sign in the image?"
**Options:** A: A yellow diamond-shaped traffic sign with a bicycle icon.
B: A hexagonal stop sign with the word "STOP" on it.
C: A rectangular traffic sign with a red border, a red circle with the number 30.
D: A road closure sign with an advance notice of 300 meters.
**Answer:** Option B

**Country:** America
**Language:** English
**Vehicle:** Car


---

**Question:** "Assume you are driving a *car* in *China*. Identify the traffic sign(s) in the image and describe the driver's required action(s). "
**Answer:** "There is a speed limit sign represented by a red circle with the text "30". This sign indicates the speed limit of the road ahead is 30 kilometers per hour, the ego car needs check the speed and keep it under the speed limit to avoid violating traffic rules."

**Country:** China
**Language:** English
**Vehicle:** Car


Figure 3: Annotated examples of three data sources – Driving Handbook Data, Driving Test Data, and Driving Road Data.

typically focus on a single question format. By including both QA and MCQ formats, IDKB enhances its utility for various applications, such as training models for open-ended and structured queries, thus providing a more comprehensive testing ground for autonomous driving models.

**Diverse Data Source** Our dataset integrates both real-world and synthetic data sources to provide a comprehensive coverage of driving scenarios. We collected driving manuals and test questions from various countries across the internet, which form the basis of our real-world data. To complement this, we enriched our dataset with data from CARLA-simulated road scenes. Relying solely on real-world road data has its limitations, as it may not cover all possible scenarios encountered in diverse driving environments. By incorporating CARLA simulations, we address this limitation and ensure that our dataset encompasses a wider range of scenarios. This combined approach allows the system to effectively translate theoretical knowledge from driving manuals and test questions into practical operational capabilities.

**Diverse Data Domain** IDKB exhibits exceptional domain diversity, covering 15 different countries, 9 languages, and 4 types of vehicles. This extensive coverage makes the dataset particularly versatile, enabling its application to various regional contexts and linguistic environments. Most existing datasets are limited to a specific country or language, typically focusing on the US and English. In contrast, IDKB's global approach ensures that models trained on it are better equipped to handle international variations in driving conditions, regulations, and vehicle types, facilitating broader applicability in autonomous driving systems worldwide.

**Diverse Knowledge Domain** Our dataset also offers comprehensive coverage of knowledge domains relevant to autonomous driving. It includes detailed information on Traffic Laws and Regulations, Road Signs and Signals, Vehicle Control and Driving Techniques, and Defensive Driving strategies. While other datasets may focus on one or two

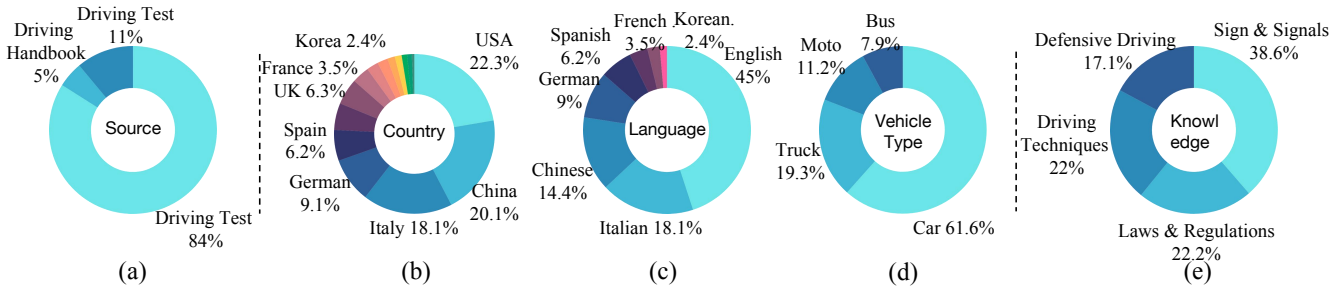



Figure 4: Data distribution in terms of data source, data domain and knowledge category.

| Model | IDKB Score | Driving Test Data | | | | | | | | | Driving Road Data | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Multi-choice Question | | | | Question & Answer | | | | Test Data Score | MCQ | QA | Road Data Score |
| | | Sing. Ans. | Mult. Ans. | Instr. Foll. | Ovl. | Rouge-1 | Rouge-L | SemScore | Ovl. | | | | |
| GPT-4o | *0.64* | *0.54* | 0.40 | *0.99* | *0.53* | *0.29* | *0.27* | *0.65* | *0.54* | *0.54* | *0.87* | 0.59 | *0.73* |
| Gemini-1.5-flash | 0.58 | 0.40 | *0.44* | *0.99* | 0.41 | 0.19 | 0.17 | 0.55 | 0.44 | 0.43 | 0.85 | *0.60* | *0.73* |
| VisualGLM-6B | 0.29 | 0.14 | 0.01 | 0.88 | 0.12 | 0.09 | 0.07 | 0.46 | 0.35 | 0.24 | 0.18 | 0.49 | 0.34 |
| Qwen-VL-chat | 0.37 | 0.12 | 0.01 | 0.61 | 0.11 | 0.13 | 0.11 | 0.50 | 0.39 | 0.25 | 0.47 | 0.50 | 0.49 |
| CogVLM | 0.39 | **0.20** | 0.03 | 0.82 | **0.18** | 0.14 | 0.12 | 0.51 | 0.40 | **0.29** | 0.40 | 0.58 | 0.49 |
| BLIP2 | 0.27 | 0.17 | 0.26 | 0.77 | **0.18** | 0.13 | 0.12 | 0.41 | 0.32 | 0.25 | 0.17 | 0.38 | 0.28 |
| LLaVA-v1.5-7B | 0.36 | 0.16 | 0.32 | 0.43 | 0.17 | 0.13 | 0.12 | 0.49 | 0.38 | 0.28 | 0.33 | 0.53 | 0.43 |
| XComposer | 0.28 | 0.14 | 0.02 | **0.99** | 0.13 | 0.13 | 0.12 | 0.46 | 0.36 | 0.25 | 0.24 | 0.37 | 0.31 |
| ShareGPT4V-7B | 0.40 | 0.17 | 0.10 | **0.99** | 0.16 | 0.17 | 0.16 | **0.52** | **0.41** | **0.29** | 0.44 | 0.58 | 0.51 |
| mPLUG-Owl2 | 0.37 | 0.18 | 0.13 | 0.59 | **0.18** | 0.14 | 0.13 | 0.51 | 0.40 | **0.29** | 0.39 | 0.51 | 0.45 |
| MiniCPM-V2.5 | 0.41 | 0.11 | 0.30 | 0.71 | 0.14 | 0.05 | 0.05 | 0.43 | 0.32 | 0.23 | 0.58 | **0.59** | 0.59 |
| XComposer2 | **0.45** | 0.11 | **0.35** | 0.91 | 0.14 | 0.15 | 0.14 | **0.52** | **0.41** | 0.28 | **0.62** | **0.59** | **0.61** |
| DeepSeek-VL-7B | 0.38 | 0.09 | 0.23 | 0.47 | 0.10 | 0.15 | 0.13 | **0.52** | **0.41** | 0.26 | 0.45 | 0.54 | 0.50 |
| Yi-VL-6B | 0.28 | 0.15 | 0.27 | 0.11 | 0.16 | 0.14 | 0.13 | 0.47 | 0.37 | 0.27 | 0.30 | 0.27 | 0.29 |
| Monkey-chat | 0.39 | 0.11 | 0.01 | 0.88 | 0.10 | **0.18** | **0.17** | 0.50 | 0.40 | 0.25 | 0.46 | 0.58 | 0.52 |

Table 2: Quantitative results for multiple LVLMs on several tasks for driving knowledge understanding. Results of proprietary LVLMs are overlaid in grey. The best results for each value of the proprietary LVLMs are highlighted in italics, while the best results for the open-source LVLMs are indicated in bold.

knowledge areas, IDKB provides a holistic view of the driving environment. This broad knowledge diversity is crucial for developing models that need to understand and navigate complex driving scenarios, making IDKB an invaluable resource for advancing autonomous driving technologies.

## 4 Benchmark

In this section, we evaluate 15 Large Vision-Language Models (LVLMs), both open-source and closed-source, using our proposed dataset. We begin by introducing the selected LVLMs and the tasks they are required to perform. Then, we outline the evaluation methods used for each task. Next, we present a quantitative evaluation of each task.

### Experiment Setup

**Selected LVLMs.** We test 15 representative LVLMs that differ in terms of parameters, open-source availability, and their vision encoders (CLIP ViT (Radford et al. 2021), EVA-CLIP-ViT (Sun et al. 2023), SAM (Kirillov et al. 2023), SigLIP (Zhai et al. 2023)) as well as their LLMs (QWen (Bai et al. 2023), Vicuna (Zheng et al. 2024), Yi (Young et al. 2024), DeepSeek (DeepSeek-AI 2024), InternLM (Cai and et al. 2024), LLaMA (Touvron et al. 2023), ChatGLM (et al. 2024), FLAN-T5 (Chung et al. 2024)). For a fair comparison, all LVLMs are used to infer questions from our dataset based on the same prompt. Further details on the selected LVLMs are provided in the Supplementary.

**Data Split.** We selected all of the driving handbook data, along with ninety percent of the driving test data and driving road data, to form the training set. The remaining ten percent of the driving test data and driving road data were used to create the test set. For more detailed statistics on the training and test sets, please refer to the Supplementary.

**Tasks.** We evaluate the LVLM's performance using two data sources: driving test data and driving road data, each containing multiple-choice questions (MCQ) and question-and-answer (QA) tasks. In the MCQ tasks, the LVLM must select the correct answer(s) from the provided options, while in the QA tasks, it is required to generate the most relevant response to a given question.

| Model | IDKB Score | Improvement | Driving Test Data | | | | Driving Road Data | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | MCQ | QA | Test Data Score | Improvement | MCQ | QA | Road Data Score | Improvement |
| GPT-4o | 0.64 | N/A | 0.53 | 0.54 | 0.54 | N/A | 0.87 | 0.60 | 0.74 | N/A |
| Qwen-VL-chat | 0.56 | 51% | 0.40 | 0.50 | 0.45 | 80% | 0.70 | 0.64 | 0.67 | 37% |
| MiniCPM-V2.5 | 0.62 | 42% | 0.42 | 0.49 | 0.46 | **100%** | **0.85** | 0.68 | **0.77** | 31% |
| XComposer2 | **0.64** | 56% | **0.51** | **0.52** | **0.52** | 86% | 0.81 | **0.71** | 0.76 | 25% |
| DeepSeek-VL-7B | 0.60 | **58%** | 0.48 | 0.51 | 0.50 | 92% | 0.68 | 0.69 | 0.69 | **38%** |

Table 3: Comparison of results for fine-tuned LVLMs. "MCQ" refers to multi-choice question and "QA" denotes question-and-answer question, "Improvements" represents the percentage increase in the three scores compared to the original value. Results of the proprietary LVLM, which has not been fine-tuned and is included for comparison, are overlaid in grey. The best results for the open-source LVLMs are indicated in bold.

## Evaluation Details

For MCQ tasks, we use regular expressions to extract options from the LVLM outputs and compare them with correct answers, measuring accuracy as the metric. Partially correct answers are not accepted in multi-answer questions. Rule-based extraction is challenging due to the free-form nature of LVLM outputs. To address this, we introduce a instruction-following test where the output must include the string "Option [A to F]" as prompted.

For QA tasks, we use ROUGE (Lin 2004) and SEMScore (Aynetdinov and Akbik 2024) to measure similarity between LVLM outputs and the reference answers. ROUGE evaluates N-gram overlap, while SEMScore assesses semantic similarity using sentence embeddings.

We calculate average scores for both MCQ and QA metrics across data sources, creating Test Data and Road Data Scores. The mean of these two scores is the IDKB Score, reflecting the LVLM's overall mastery of driving knowledge. More details about metrics are provided in Supplementary.

## Main Results

In this subsection, we analyze various LVLMs on our test set, focusing on overall performance, distinctions between driving test and road data and comparison between proprietary and open-source LVLMs. We also highlight the impact of fine-tuning with our dataset, showcasing key improvements in model performance. More details and analysis will be provided in Supplementary.

**Overall Evaluation** In Tab. 2, we present the performance results of various VLMs on our test set. GPT-4o achieved the highest overall performance with an IDKB score of 0.64. Among the open-source models, XComposer2 stood out with the best performance, achieving an IDKB score of 0.45. Most open-source models fell within an IDKB score range from 0.35 to 0.4. However, BLIP2, XComposer, Yi-VL and VisualGLM underperformed, with IDKB scores of 0.27, 0.28, 0.28 and 0.29, respectively. Overall, the evaluated LVLMs generally did not demonstrate strong driving domain knowledge, highlighting the need for high-quality, structured and diverse driving knowledge data for effective applications in autonomous driving.

**Driving Test Data vs. Driving Road Data** LVLMs generally perform better on driving road data than on driving test data. Most open-source LVLMs scored around 0.25 on Test Data Score, while the average Road Data Score was 0.44. A similar trend was observed in the two proprietary LVLMs. This suggests that while many LVLMs have a basic understanding of traffic signs, they lack a deeper comprehension of traffic laws, regulations, and driving skills—areas more thoroughly assessed in driving test data.

**Proprietary vs. Open-Source** Proprietary LVLMs generally outperform open-source LVLMs in evaluation results, likely due to their larger number of parameters and more extensive knowledge base. This advantage is more pronounced with driving test data, which requires external knowledge of laws and rules—areas where proprietary models excel. In contrast, when dealing with driving road data, which emphasizes traffic sign recognition, some open-source LVLMs like XComposer2 can achieve performance levels comparable to those of proprietary models.

**Significant Improvement through Fine-Tuning** To evaluate the impact of structured driving knowledge data, we fine-tuned four representative LVLMs with different visual encoders or LLMs. As shown in Table 3, the fine-tuned LVLMs achieved IDKB scores comparable to or matching those of much larger proprietary models. MiniCPM-V2.5's Test Data Score doubled, indicating our data enhances expertise in driving laws, rules, techniques, and special scenarios. Improved Test Road Scores also suggest better traffic sign interpretation and understanding of road regulations. These results underline our dataset's role in advancing LVLM competence, contributing to safer, more reliable autonomous driving systems.

# 5 Conclusion

In this paper, we introduced IDKB, a pioneering large-scale dataset designed to bridge the gap in domain-specific driving knowledge within LVLMs. IDKB includes over 1 million entries on driving regulations, scenarios, and practices from 15 countries and 9 languages. We evaluated 15 LVLMs using IDKB to assess their performance in the context of autonomous driving and provided extensive analysis. Additionally, we fine-tuned several popular models, achieving significant improvements in their performance, which further highlights the importance of our dataset.

## Acknowledgements

This work was supported by NSFC (No.62206173), Shanghai Frontiers Science Center of Human-centered Artificial Intelligence (ShangHAI), MoE Key Laboratory of Intelligent Perception and Human-Machine Collaboration (KLIP-HuMaCo).

## References

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Aynetdinov, A.; and Akbik, A. 2024. SemScore: Automated Evaluation of Instruction-Tuned LLMs based on Semantic Textual Similarity. arXiv:2401.17072.

Bai, J.; Bai, S.; Yang, S.; Wang, S.; Tan, S.; Wang, P.; Lin, J.; Zhou, C.; and Zhou, J. 2023. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*.

Cai, Z.; and et al., M. C. 2024. InternLM2 Technical Report. arXiv:2403.17297.

Chung, H. W.; Hou, L.; Longpre, S.; Zoph, B.; Tay, Y.; Fedus, W.; Li, Y.; Wang, X.; Dehghani, M.; Brahma, S.; et al. 2024. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70): 1–53.

DeepSeek-AI. 2024. DeepSeek LLM: Scaling Open-Source Language Models with Longtermism. *arXiv preprint arXiv:2401.02954*.

Deruyttere, T.; Vandenhende, S.; Grujicic, D.; Van Gool, L.; and Moens, M.-F. 2019. Talk2car: Taking control of your self-driving car. *arXiv preprint arXiv:1909.10838*.

Dosovitskiy, A.; Ros, G.; Codevilla, F.; Lopez, A.; and Koltun, V. 2017. CARLA: An open urban driving simulator. In *Conference on robot learning*, 1–16. PMLR.

et al., T. G. 2024. ChatGLM: A Family of Large Language Models from GLM-130B to GLM-4 All Tools. arXiv:2406.12793.

Jia, X.; Yang, Z.; Li, Q.; Zhang, Z.; and Yan, J. 2024. Bench2Drive: Towards Multi-Ability Benchmarking of Closed-Loop End-To-End Autonomous Driving. *arXiv preprint arXiv:2406.03877*.

Kim, J.; Rohrbach, A.; Darrell, T.; Canny, J.; and Akata, Z. 2018. Textual explanations for self-driving vehicles. In *Proceedings of the European conference on computer vision (ECCV)*, 563–578.

Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A. C.; Lo, W.-Y.; et al. 2023. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4015–4026.

Li, K.; Chen, K.; Wang, H.; Hong, L.; Ye, C.; Han, J.; Chen, Y.; Zhang, W.; Xu, C.; Yeung, D.-Y.; et al. 2022. Coda: A real-world road corner case dataset for object detection in autonomous driving. In *European Conference on Computer Vision*, 406–423. Springer.

Li, Y.; Zhang, W.; Chen, K.; Liu, Y.; Li, P.; Gao, R.; Hong, L.; Tian, M.; Zhao, X.; Li, Z.; et al. 2024. Automated evaluation of large vision-language models on self-driving corner cases. *arXiv preprint arXiv:2404.10595*.

Lin, C.-Y. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, 74–81.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2024. Visual instruction tuning. *Advances in neural information processing systems*, 36.

Malla, S.; Choi, C.; Dwivedi, I.; Choi, J. H.; and Li, J. 2023. Drama: Joint risk localization and captioning in driving. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 1043–1052.

Mao, J.; Ye, J.; Qian, Y.; Pavone, M.; and Wang, Y. 2023. A language agent for autonomous driving. *arXiv preprint arXiv:2311.10813*.

Park, S.; Lee, M.; Kang, J.; Choi, H.; Park, Y.; Cho, J.; Lee, A.; and Kim, D. 2024. Vlaad: Vision and language assistant for autonomous driving. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 980–987.

Qian, T.; Chen, J.; Zhuo, L.; Jiao, Y.; and Jiang, Y.-G. 2024. Nuscenes-qa: A multi-modal visual question answering benchmark for autonomous driving scenario. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 4542–4550.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Shao, H.; Hu, Y.; Wang, L.; Song, G.; Waslander, S. L.; Liu, Y.; and Li, H. 2024. Lmdrive: Closed-loop end-to-end driving with large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15120–15130.

Sima, C.; Renz, K.; Chitta, K.; Chen, L.; Zhang, H.; Xie, C.; Luo, P.; Geiger, A.; and Li, H. 2023. Drivelm: Driving with graph visual question answering. *arXiv preprint arXiv:2312.14150*.

Sun, Q.; Fang, Y.; Wu, L.; Wang, X.; and Cao, Y. 2023. Eva-clip: Improved training techniques for clip at scale. *arXiv preprint arXiv:2303.15389*.

Tian, X.; Gu, J.; Li, B.; Liu, Y.; Hu, C.; Wang, Y.; Zhan, K.; Jia, P.; Lang, X.; and Zhao, H. 2024. Drivevlm: The convergence of autonomous driving and large vision-language models. *arXiv preprint arXiv:2402.12289*.

Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Xu, Y.; Yang, X.; Gong, L.; Lin, H.-C.; Wu, T.-Y.; Li, Y.; and Vasconcelos, N. 2020. Explainable object-induced action decision for autonomous vehicles. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9523–9532.

Xu, Z.; Zhang, Y.; Xie, E.; Zhao, Z.; Guo, Y.; Wong, K. K.; Li, Z.; and Zhao, H. 2023. Drivegpt4: Interpretable end-to-end autonomous driving via large language model. *arXiv preprint arXiv:2310.01412*.

Young, A.; Chen, B.; Li, C.; Huang, C.; Zhang, G.; Zhang, G.; Li, H.; Zhu, J.; Chen, J.; Chang, J.; et al. 2024. Yi: Open foundation models by 01. ai. *arXiv preprint arXiv:2403.04652*.

Zhai, X.; Mustafa, B.; Kolesnikov, A.; and Beyer, L. 2023. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 11975–11986.

Zheng, L.; Chiang, W.-L.; Sheng, Y.; Zhuang, S.; Wu, Z.; Zhuang, Y.; Lin, Z.; Li, Z.; Li, D.; Xing, E.; et al. 2024. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36.