

# DriveMM: All-in-One Large Multimodal Model for Autonomous Driving

Zhijian Huang<sup>1†</sup> Chengjian Feng<sup>2†</sup> Feng Yan<sup>2</sup> Baihui Xiao<sup>2</sup> Zequn Jie<sup>2</sup>  
 Yujie Zhong<sup>2</sup> Xiaodan Liang<sup>1‡</sup> Lin Ma<sup>2‡</sup>

<sup>1</sup>Shenzhen Campus of Sun Yat-sen University <sup>2</sup>Meituan Inc.

<https://zhijian11.github.io/DriveMM>

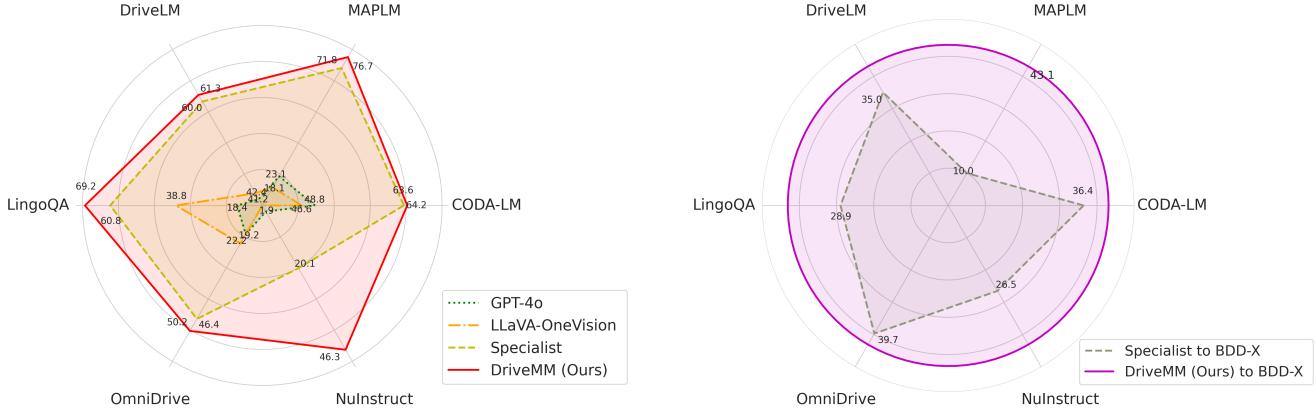


Figure 1. **DriveMM achieves SOTA in both general capabilities and generalization ability.** **Left:** DriveMM outperforms all specific SOTA models and other general large multimodal models across all 6 datasets comprising 13 tasks; **Right:** In zero-shot learning on unseen dataset [17], DriveMM demonstrates stronger generalization ability compared to specialist models trained on individual datasets.

## Abstract

Large Multimodal Models (LMMs) have demonstrated exceptional comprehension and interpretation capabilities in Autonomous Driving (AD) by incorporating large language models. Despite the advancements, current data-driven AD approaches tend to concentrate on a single dataset and specific tasks, neglecting their overall capabilities and ability to generalize. To bridge these gaps, we propose DriveMM, a general large multimodal model designed to process diverse data inputs, such as images and multi-view videos, while performing a broad spectrum of AD tasks, including perception, prediction, and planning. Initially, the model undergoes curriculum pre-training to process varied visual signals and perform basic visual comprehension and perception tasks. Subsequently, we augment and standardize various AD-related datasets to fine-tune the model, resulting in an all-in-one LMM for autonomous driving. To assess the general capabilities and generalization ability, we conduct evaluations on six public benchmarks and undertake zero-shot transfer on an unseen dataset, where DriveMM achieves state-of-the-art performance across all tasks. We hope DriveMM as a promising solution for future end-to-end autonomous driving applications in the real world.

† Equal contribution. ‡ Corresponding author.

## 1. Introduction

Recently, vision-language driving datasets and models have garnered significant attention in the field of autonomous driving [6, 54, 58]. Numerous datasets [4, 13, 17, 22, 34, 45, 49] have been meticulously developed and curated to fine-tune Large Multimodal Models (LMMs), enabling them to better understand and generate multimodal content, as well as adapt to specific domains and applications. Meanwhile, several methods [11, 15, 32, 33, 38, 42, 48, 50, 51, 53] have attempted to incorporate the extensive world knowledge and strong logical reasoning capabilities of Large Language Model (LLM) into AD systems, demonstrating significant improvements in interpretability and overall system performance. Typically, these methods utilize a pre-training and fine-tuning approach [26], where LMMs are fine-tuned on certain datasets and evaluated on the tasks within.

Due to the complexity [12, 27] and diversity [14, 56] of driving scenarios and driver behaviors [10, 36], existing LMMs and AD datasets often focus on specific scenes and tasks, as shown in Tab. 1. We observe that the methodologies employed in the collection of various datasets are tailored to the specific tasks they are designed to address. For instance, the CODA-LM dataset [22], which is centered on corner case perception, necessitates only front view images.

Dataset	Type	Perception									Prediction		Planning		
		Scene Und.	Region Und.	Key Und.	Corner Und.	Road Und.	Risk Det.	Key Det.	Key Gro.	Status Pre.	Motion Pre.	Action Dec.	Driving Res.	Motion Pre.	
CODA-LM[22]	S.I.	X	✓	X	✓	X	X	X	X	X	X	✓	✓	X	
MAPLM[4]	M.I.	X	X	X	X	✓	X	X	X	X	X	X	X	X	
DriveLM[45]	M.I.	✓	✓	✓	X	X	X	✓	✓	✓	✓	✓	✓	X	
LingoQA[34]	S.V.	✓	✓	X	X	X	X	X	X	X	X	✓	✓	X	
OmniDrive[49]	M.V.	✓	✓	✓	X	X	X	X	✓	X	X	✓	✓	✓	
NuInstruct[13]	M.V.	X	X	X	X	✓	✓	✓	✓	✓	✓	X	✓	✓	

Table 1. **Comparison of different AD datasets.** Different datasets encompass various input types and multiple sub-tasks of perception, prediction, and planning in real-world scenarios. S.I.=Single-view image, M.I.=Multi-view images, S.V.=Single-view video, M.V.=Multi-view videos. Und.=Understanding, Det.=Detection, Gro.=Grounding, Pre.=Prediction, Dec.=Decision, Res.=Reasoning.

In contrast, the NuInstruct dataset which is derived from the nuScenes dataset [3], encompasses tasks related to prediction and decision-making, necessitating multi-view or video inputs. Furthermore, each dataset is annotated to address distinct sub-tasks; for example, MapLM [4] is dedicated to road-related perception tasks, whereas LingoQA [34] is oriented towards planning tasks, emphasizing the action decision and driving reasoning of the ego vehicle. As illustrated in Fig. 1, previous specialist LMMs [4, 13, 21, 22, 34, 49] fine-tuned on these single datasets lack the general capability needed to handle the complex and varied tasks found in real-world scenarios (Left) and exhibit poor generalization performance on another dataset (Right). This limitation highlights the need for a more general LMM to improve the versatility and robustness of AD tasks.

Therefore, we motivate an all-in-one general-purpose LMM, which is capable of simultaneously accepting various types of data inputs (*e.g.*, images, videos) and performing a wide range of tasks in AD (*e.g.*, perception, prediction, planning). To begin with, we re-engineer an LMM to accept perspective-aware visual signals, by providing an explanation about the camera perspective and data type in the instruction. It allows the model to recognize the spatial relationships of objects and analyze the full context of dynamic driving environments. To effectively train the all-in-one LMM, we employ a curriculum principle [28] to pre-train and fine-tune the model. This approach gradually guides the model to handle intricate data inputs, progressing from single image to multi-view videos, as well as diverse tasks, transitioning from image captioning to driving reasoning. In the pre-training phase, we begin by equipping LLM with a foundational ability to comprehend images through training it on image-text pairs. To obtain a robust foundational LMM, we further conduct multi-capability pre-training, leveraging diverse types of multimodal and perception data to enhance the model’s visual reasoning and perception capacities in different scenarios. During the fine-tuning stage, we gather various open-source multimodal AD datasets presented in Tab. 1, and then enhance

their mutual improvement by augmenting and standardizing their question-answer pairs. By integrating and leveraging these diverse data and tasks, our model, **DriveMM**, is capable of efficiently performing various AD tasks in real-world scenarios. Experimentally, we thoroughly evaluate our DriveMM on all challenging benchmarks, where it achieves state-of-the-art performance across all tasks.

To summarize, our contributions are:

- We propose a novel all-in-one large multimodal model, DriveMM, robustly equipped with the general capabilities to execute a wide range of AD tasks and the generalization ability to effectively transfer to new datasets.
- We introduce comprehensive benchmarks for evaluating autonomous driving LMMs, which include six public datasets, four input types, and thirteen challenging tasks. To the best of our knowledge, this is the first to use multiple benchmarks to evaluate autonomous driving LMMs.
- We present a curriculum principle for pre-training and fine-tuning on both diverse multimodal data and AD data. DriveMM demonstrates state-of-the-art performances and consistently outperforms models trained on the individual dataset across all evaluated benchmarks.

## 2. Related Work

### 2.1. Vision-Language Driving Datasets

In recent years, numerous vision-language driving datasets have been developed with the aim of training and evaluating LMMs designed for AD scenarios. [17, 39, 52, 53] are dedicated to scene description and actions decision in driving videos. DRAMA [30], CODA-LM [22], and DriveVLM [46] focus on risk objects and corner cases learning. In addition to the single-view data, many studies construct multi-view data based on the nuScenes dataset [3]. For instance, NuScenes-QA [40] introduces free-form question-answer annotations for 3D object relationships. DriveLM [45], OmniDrive [49], and NuInstruct [13] employ the original annotations and LLMs to generate visual question-answer pairs covering perception,

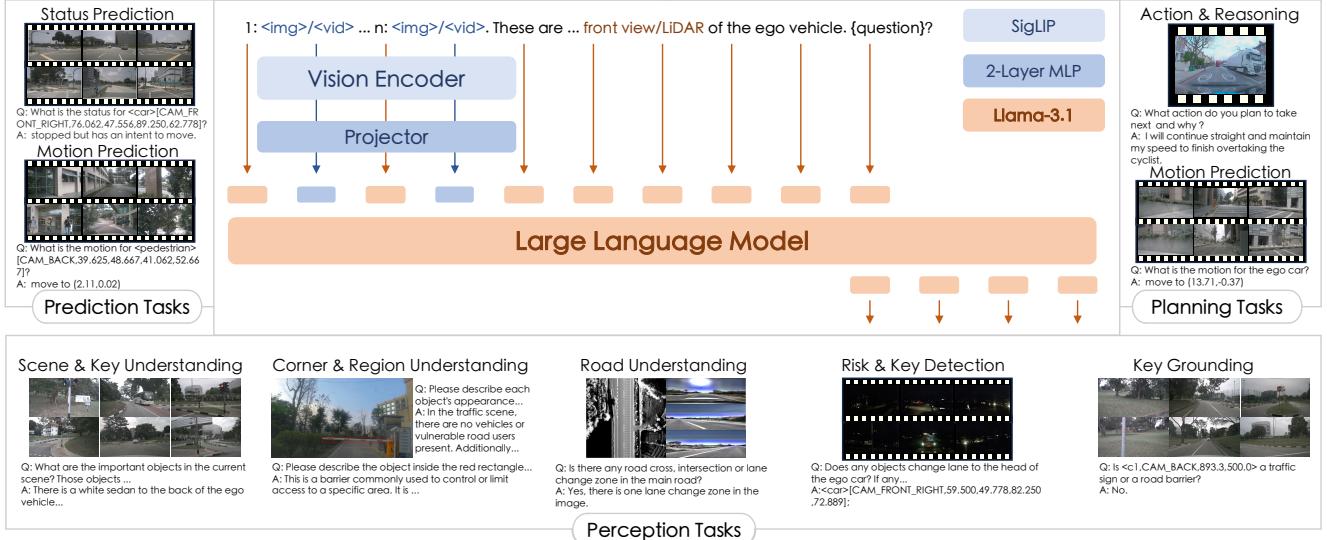


Figure 2. **Overview of DriveMM framework.** We adapt the architecture form of LLava [26] with a different model instantiation, processing various visual input signals. We design a perspective-aware prompt to accept multi-perspective inputs in AD scenario. Equipped with diverse AD multimodal data, DriveMM possesses an all-in-one capability to accomplish multiple tasks in autonomous driving.

reasoning, and planning. Furthermore, MAPLM [4] integrates multi-view data and LiDAR data to analyze and recognize road surface conditions. In this paper, we augment and standardize multiple driving datasets to train a comprehensive LMM for diverse autonomous driving scenarios.

## 2.2. LMMs for Autonomous Driving

LMMs have demonstrated impressive performance in diverse tasks [1, 7, 24, 26, 29]. Recently, researchers have begun to explore the potential of LLMs in the field of AD. In the early stages, DiLu [51] and GPT-Driver [31] attempt to utilize GPT-3.5 and GPT-4 as driving planners. Subsequently, DriveGPT4 [53] and RDA-Driver [15] introduce end-to-end LMMs that generate control signals or trajectories. Unlike the methods that handle driving maneuvers through language, LMDrive [42] and DriveMLM [50] use a decoder to predict control signal from hidden embeddings. In order to enhance the perception and reasoning abilities, several approaches aim to improve the model architecture. Reason2Drive [37] proposes a prior tokenizer to extract local image features and BEV-InMLLM [13] injects Bird's-Eye-View (BEV) representations into LMMs. OmniDrive [49] uses Q-Former3D to integrate 2D pre-trained knowledge with essential 3D spatial understanding. ELM [59] incorporates a time-aware token selection module to accurately inquire about temporal cues. Although these methods have demonstrated satisfactory performance, their applicability is limited to the specific scene and task, such as a particular data type or a dataset-specific task. In light of this, we propose an all-in-one LMM designed to efficiently process diverse driving scenes and tasks in AD.

## 3. Methodology

### 3.1. Overview

In this paper, we propose DriveMM, an all-in-one LMM designed to efficiently process various driving data and tasks in AD. Formally, given a visual signal  $X_v$  captured by the vehicle sensors and a user instruction  $X_t$ , DriveMM  $\mathcal{F}(\cdot)$  provides the driving-related analysis and suggestions:

$$Y_t = \mathcal{F}(X_v, X_t). \quad (1)$$

$X_v$  can represent various data formats, including image, multi-images, video, and multi-videos captured by a single camera, multi-view cameras, or LiDAR, while  $X_t$  encompasses questions pertaining to perception, prediction, reasoning, decision-making, and more. By integrating diverse data and tasks, *DriveMM can be trained on a wide range of AD vision-language data, resulting in mutual improvements across different datasets and tasks*. Moreover, once trained, *DriveMM can be effectively deployed across a broad spectrum of real-world AD scenarios, e.g. different camera and radar system configurations, as well as various AD tasks*.

In the following sections, we first describe the architecture of DriveMM, which has the capability to process multiple types of data captured by different sensors (Sec. 3.2). To facilitate the model's comprehension of AD scenarios, we gather diverse datasets with multiple data formats and tasks, then augment and standardize their question-answer pairs to enhance collaboration across different datasets (Sec. 3.3). In order to effectively train DriveMM on various datasets and tasks, we adopt a curriculum learning approach to progressively enhance the model's capability (Sec. 3.4).

### 3.2. Model Architecture

Our goal is to design an efficient model architecture that can synchronously tackle single image, multi-images, single-view video, and multi-view video in AD scenarios. Shown in Fig. 2, DriveMM follows a design of the predominant LMMs like LLaVA [26]. It comprises three components: a vision encoder  $\mathcal{F}_e(\cdot)$ , a projector  $\mathcal{F}_p(\cdot)$  and a LLM  $\mathcal{F}_l(\cdot)$ .

In particular, the vision encoder (*e.g.* SigLIP [55] in this paper) encodes the input images  $X_v \in \mathbb{R}^{(n \times f) \times h \times w \times 3}$  into the visual features:

$$Z_v = \mathcal{F}_e(X_v), \quad (2)$$

where  $Z_v \in \mathbb{R}^{(n \times f) \times h' \times w' \times d'}$ ,  $n$  and  $f$  denote the number of cameras and frames,  $(h, w)$  and  $(h', w')$  denote the size of image and feature, and  $d'$  denotes the channel dimensionality.  $X_v$  can represent the data formats mentioned above, for example,  $n > 1$  and  $f > 1$  for video from multi-view cameras. For LiDAR data, we project the point clouds onto the BEV or range view to convert the data into a single image format. Afterward, the projector projects the image features into the word embedding space:

$$H_v = \mathcal{F}_p(Z_v), \quad (3)$$

where the projector is implemented using a 2-layer MLP,  $H_v \in \mathbb{R}^{(n \times f \times h \times w) \times d}$  denotes a sequence of visual tokens, and  $d$  denotes the dimensionality of the word embedding space in LLM (*e.g.* LLaMA-3.1 [47]). Based on the visual tokens  $H_v$  and the user instruction  $X_t$ , the LLM computes the probability of the target word step by step:

$$p(Y_t|H_v, X_t) = \prod_{i=1}^L \mathcal{F}_l(Y_{t,i}|H_v, \Phi(X_t), Y_{t,0:i-1}), \quad (4)$$

where  $\Phi(\cdot)$  refers the text tokenizer,  $Y_{t,i}$  and  $Y_{t,0:i-1}$  represent the  $i$ th word and the preceding  $i - 1$  words in  $Y_t$ , and  $L$  indicates the length of the words generated by LLM.

**Perspective-aware prompt.** In Eq. (4), typical LMMs [2, 26] flatten visual features for LLM input, failing to distinguish between perspectives (*e.g.*, front or back view) and formats (*e.g.*, image or video). To address this, we propose a perspective-aware prompt. As shown in the Tab. 2, we utilize different placeholders (*i.e.* `<image>` and `<video>`) for image and video inputs, where the placeholders will be replaced by respective tokens before being fed into LLM. We also assign numerical labels to images/videos with different perspectives and explain the specific camera or LiDAR for each in the text. In order to enhance computational efficiency, we apply a  $2 \times 2$  spatial pooling on the video features  $H_v$ , and then flatten them into the visual tokens. Incorporating the information of perspective and data format, DriveMM can better interpret complex traffic situations, recognize multiple objects and their spatial relationships, and make more informed decisions.

#### Perspective-aware Prompt:

1: `<image>/<video>` 2: `<image>/<video>` ... n: `<image>/<video>`. These n images/videos are the front view, front left view, ..., and LiDAR of the ego vehicle. {question}?

Table 2. The perspective-aware prompt for multi-view inputs.

### 3.3. Data

In the training of LMMs, data plays a crucial role in enabling and activating the LLM’s ability to understand multimodal information. To enhance the comprehension and reasoning skills of DriveMM in multimodal AD scenarios, we construct three distinct datasets: conventional multimodal data, perception data, and autonomous driving data.

#### 3.3.1. Conventional Multimodal Data

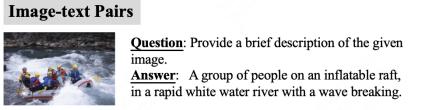
Recent studies [2, 9] show that LMMs can achieve enhanced performance as the volume of data increases. However, compared to the abundant image-text data available online [41, 44], AD image-text data is significantly limited. To enhance the performance of DriveMM, we pre-train a base model with extensive multimodal data, enabling reasoning with single images, multi-images, and video.

Specifically, we construct a multimodal dataset from [18] comprising image-text pairs and diverse visual instruction tuning data. The objective of the image-text pairs is to align the vision encoder and LLM, enabling the model to develop a foundational understanding of images. We utilize multiple datasets, including LCS-558K [26], COCO118K [18], CC3M [44]. To enhance the model’s capability in dealing with the visual data in various sensor configurations such as single-view and multi-view cameras, we utilize visual instruction tuning data in the OneVision data [19], including image, multi-images and video.

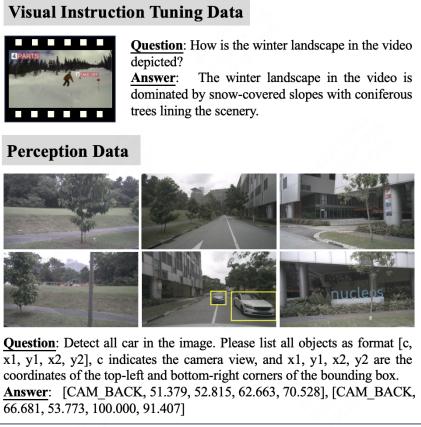
#### 3.3.2. Perception Data

To equip DriveMM with AD perception capabilities, we create a comprehensive grounding dataset including various data formats. For the single image data, we utilize the COCO [23] and Object365 [43] datasets. We randomly select a category from an image and use grounding prompts (*e.g.*, “Detect all `<category>` in the image.”) to prompt the model to detect all objects in that category. We represent the object’s position with the bounding box  $[x_{min}, y_{min}, x_{max}, y_{max}]$  or the region center  $[x_{center}, y_{center}]$ . The  $x$  and  $y$  values are normalized in a range of 0 to 100 based on the image’s size. For the multi-view images and multi-view videos, we employ the nuScenes [3] dataset. To imbue the model with a sense of spatial awareness, we expect it not only to predict the object bounding boxes but also to estimate the camera perspective. Therefore, we represent the object’s position with  $[cam, x_{min}, y_{min}, x_{max}, y_{max}]$  or

## Stage-1 & Stage-2: Image pre-training



## Stage-3: Multi-capacity pre-training



## Stage-4: Driving fine-tuning

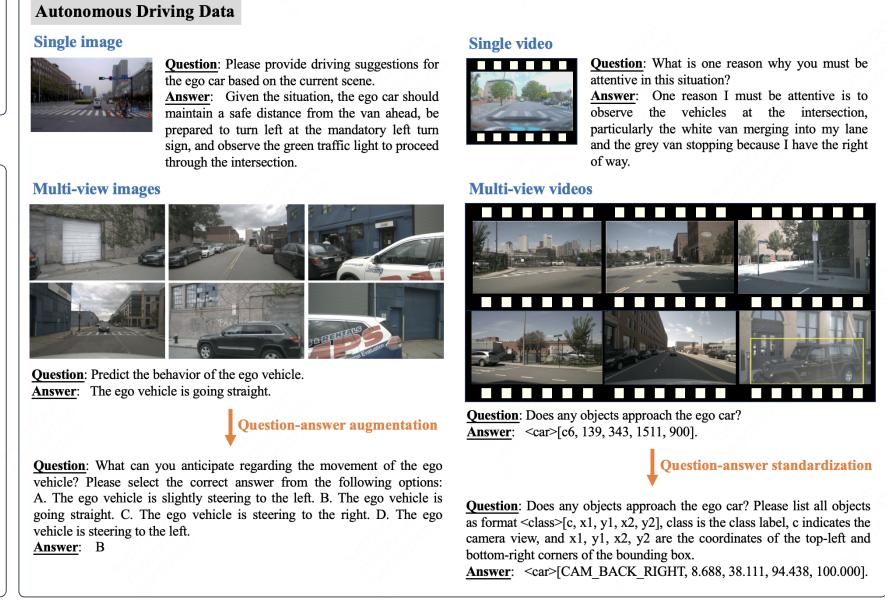


Figure 3. Illustration of the curriculum learning framework. **Stage-1 & Stage-2:** it consists of language-image alignment and single-image pre-training, which use the image-text pairs to equip LLM with a foundational capability for single-image comprehension. We refer to the combination of these two stages as image pre-training. **Stage-3:** we enhance the model’s visual reasoning and perception capabilities across diverse scenarios by training on both the visual instruction tuning data and perception data. **Stage-4:** we further fine-tune the model on six augmented and standardized autonomous driving datasets, enabling DriveMM to tackle a wide range of autonomous driving tasks.

[cam, x<sub>center</sub>, y<sub>center</sub>], where cam denotes the camera perspective such as “CAM\_BACK”. An example of the perception data is illustrated in the bottom left of Fig. 3.

### 3.3.3. Autonomous Driving Data

Here we collect diverse datasets to train an all-in-one LMM that can synchronously tackle various AD tasks in different scenarios. Concretely, we use six autonomous driving datasets: CODA-LM [22], MAPLM [4], DriveLM [45], LingoQA [34], OmniDrive [49] and NuInstruct [13]. The detailed descriptions of the six datasets are shown in Tab. 1. These datasets encompass various sensor configurations such as camera and LiDAR, and different AD tasks including perception, prediction, and planning. It is noteworthy to mention that different datasets may exhibit distinct question patterns. To foster collaborative enhancement, we augment and standardize the question-answer pairs as follows.

**Question-answer augmentation.** Some datasets are restricted to a fixed set of templates. For instance, CODA-LM comprises only three question templates, while MAPLM utilizes five. It hinders the potential for the model’s generalization. To overcome this limitation, we employ GPT-4o-mini to augment the question-answer pairs and increase their diversity. Additionally, a significant portion of the questions are open-ended. To further enhance the diversity, we randomly transform some open-ended questions into multiple-choice format. An example of the augmenta-

tion is illustrated in the bottom right of Fig. 3. Please refer to the supplementary material for more details.

**Question-answer standardization.** Different datasets may exhibit inconsistencies in question-answer styles. For example, DriveLM uses “<c6, CAM\_BACK, 1088.3, 497.5>” to represent an object, where “c6” denotes the class ID. In contrast, NuInstruct employs the format of “<car>[c6, 139, 343, 1511, 900]”, where “c6” represents the camera ID. To ensure compatibility across datasets, we standardize the representation of objects and explicitly specify the representation format. Moreover, to accommodate bounding boxes in images of different sizes, we standardize the coordinates of the bounding boxes to a range of 0 to 100 based on the image’s size. For example, for the NuInstruct dataset, we re-represent the object as “<car>[CAM\_BACK\_RIGHT, 8.688, 38.111, 94.438, 100.000]” and add the formatting instructions at the end of the question, as illustrated in the bottom right of Fig. 3.

## 3.4. Training

In this section, we present a curriculum learning approach to progressively improve the performance of the model on various AD data and tasks, resulting in an all-in-one autonomous driving model DriveMM. Specifically, we gradually increase the complexity of the data, progressing from a single image to multiple videos, and the complexity of the

Dataset	Metrics	Specialist Model <sup>†</sup>	Generalist Model			
			GPT-4o [16]	LLaVA-OV [19]	Drive-OV	DriveMM
CODA-LM [22]	General↑	<b>55.04</b>	47.06	38.70	53.60	52.94
	Regional↑	77.68	50.37	51.70	77.64	<b>77.76</b>
	Suggestion↑	58.14	48.94	49.32	58.72	<b>61.84</b>
	Average ↑	63.62	48.79	46.57	63.32	<b>64.18</b>
MAPLM [4]	FRM↑	57.99	0.33	0.00	62.33	<b>64.80</b>
	QNS↑	85.52	45.9	36.18	87.55	<b>88.53</b>
	Average↑	71.76	23.12	18.09	74.94	<b>76.67</b>
DriveLM [45]	Accuracy↑	73.39	38.55	25.03	<b>79.38</b>	76.09
	ChatGPT↑	65.25	67.27	65.70	65.92	<b>66.44</b>
	Language↑	48.56	8.97	14.44	47.28	<b>48.90</b>
	Match↑	47.65	24.00	40.93	40.70	<b>48.63</b>
	Average ↑	60.02	41.21	42.36	59.84	<b>61.30</b>
LingoQA [34]	Lingo-Judge↑	60.80	18.40	38.80	<b>70.10</b>	69.20
OmniDrive [49]	BLEU↑	38.00	10.91	16.14	38.25	<b>39.11</b>
	CIDEr↑	68.60	24.42	28.41	76.04	<b>77.50</b>
	ROUGE↑	32.60	22.34	22.14	33.36	<b>34.15</b>
	Average ↑	46.40	19.22	22.23	49.22	<b>50.25</b>
NuInstruct [13]	MAE↓	9.08	9.93	87.04	1.81	<b>1.56</b>
	Accuracy↑	32.48	10.64	3.75	64.57	<b>64.71</b>
	MAP↑	21.93	0.00	0.00	23.39	<b>39.04</b>
	BLEU↑	35.20	7.08	8.55	80.85	<b>83.00</b>
	Average* ↑	20.13	1.95	0.00	41.75	<b>46.30</b>

Table 3. General performance on benchmarks. We compare with state-of-the-art specialist models, commercial models and open-source large multimodal models across diverse autonomous driving valuation benchmarks spanning multiple modalities. <sup>†</sup>Specialist models correspond to the performance of six different models [4, 13, 21, 22, 34, 49]. \* indicates  $\max((\text{Accuracy} + \text{MAP} + \text{BLEU} - \text{MAE}) / 4, 0)$ .

tasks, transitioning from image captioning to driving reasoning, for training DriveMM. As illustrated in Fig. 3, the training process is divided into four steps:

**Stage-1: Language-image alignment.** The goal of this stage is to equip the pre-trained LLM with a foundational capability for multimodal comprehension. To achieve that, we train the projector to align with the word embedding space of LLM. We froze both the vision encoder and LLM, and only optimize the projector on LCS-558K [26].

**Stage-2: Single-image pre-training.** In this stage, we further enhance the model’s capacity to comprehend single-image by collectively optimizing the entire model. We use the image-text pairs outlined in Sec. 3.3.1, and optimize all the parameters of the model to enhance the suitability of LLM for multimodal tasks.

**Stage-3: Multi-capacity pre-training.** To obtain a robust foundational model for training AD systems, we enhance the model’s reasoning and perception capabilities across diverse scenarios. To this end, we utilize the visual instruction tuning data described in Sec. 3.3.1 to enhance the model to reason about fundamental visual elements. Additionally, we employ the perception data described in Sec. 3.3.2 to facilitate the model’s perception capacity. It is noteworthy that

the training data comprises diverse data formats, including single image, single video, multi-view images, and multi-view videos. By equipping the model with the capability to process various data and tasks, we establish the groundwork for training an all-in-one AD model.

**Stage-4: Driving fine-tuning.** To enable DriveMM to tackle a wide range of AD tasks, we further fine-tune the model on diverse driving datasets. Specifically, we utilize six augmented and standardized autonomous driving datasets outlined in Sec. 3.3.3. In this stage, we optimize all the parameters of the model. Once trained, the proposed all-in-one DriveMM can be effectively deployed across a broad spectrum of AD scenarios, *e.g.* different camera and radar system configurations, as well as various AD tasks.

## 4. Experiment

### 4.1. Experimental Setting

#### 4.1.1. Dataset

We utilize six open-source autonomous driving datasets: CODA-LM [22], MAPLM [4], DriveLM [45], LingoQA [34], OmniDrive [49] and NuInstruct [13]. DriveLM, OmniDrive and Nuinstruct are annotated based on

Dataset	Metrics	Specialist Model w/						Generalist Model
		CODA-LM	MAPLM	DriveLM	LingoQA	OmniDrive	NuInstruct	
BDD-X[17]	GPT-Score	36.40	10.01	35.04	28.93	39.67	26.48	43.10

Table 4. Generalization ability in BDD-X. Specialists are fine-tuned on a single dataset, whereas DriveMM is fine-tuned on all datasets.

the nuScenes [3] dataset and contain 376,181, 374,329 and 71,842 samples respectively, suitable for multi-view images or videos input. The CODA-LM dataset, designed for corner case question-answer pairs on the CODA-LM [22] dataset, includes a total of 184,480 samples in both Chinese and English, as a single-view dataset. MAPLM, a multi-view image dataset, contains 94,970 samples, while LingoQA, a single-view video dataset, comprises 413,829 samples. Note the sample number is computed after our enhancement. Additionally, each dataset includes a specific number of test samples: DriveLM has 15,480 test samples, OmniDrive has 72,184 test samples, NuInstruct has 16,147 test samples, CODA-LM has 2,123 test samples, MAPLM has 6,642 test samples, and LingoQA has 500 test samples.

#### 4.1.2. Evaluation Metrics

We follow the common practice metrics in each work for fair comparison: (1) CODA-LM uses text-only LLMs, *e.g.* GPT-4, as evaluators to score model responses by few-shot learning. (2) MAPLM evaluates fine-grained QAs with multi-class classification correct ratio as the accuracy metric, while open QAs are assessed with a rule-based BLEU metric. (3) DriveLM implements four evaluation methods: accuracy, LLM score, language rule-based evaluation, and match score. (4) LingoQA uses a learned text classifier Lingo-Judge to estimate the score of model answers. (5) OmniDrive employs rule-based language metrics to evaluate sentence similarity at the word level. (6) NuInstruct uses a variety of metrics to evaluate different tasks: Mean Absolute Error (MAE) for regression tasks, accuracy for classification tasks, Mean Average Precision (MAP) for detection tasks, and a rule-based BLEU metric for captioning tasks.

## 4.2. Main Results

We compare DriveMM with current state-of-the-art models and report the result in all six benchmarks. As depicted in Tab. 3, DriveMM surpasses the previous works across all benchmarks, such as a remarkable increase of **+4.91** in MAPLM and **+26.17** in NuInstruct. It is worth noting that the specialist LLMs (*e.g.* CODA-VLM[22], InternVL4Drive-v2[21]) trained on specific datasets have shown excellent performance on specific tasks. However, by utilizing extensive multi-type information from multiple real-world AD datasets, DriveMM not only simultaneously accomplishes all tasks in real scenarios but also outperforms all specialist models. This highlights our model’s ability

to deliver an all-in-one solution by effectively leveraging a wide range of autonomous driving data and tasks.

Furthermore, DriveMM demonstrates superior performance compared to GPT-4o [16]. This suggests that our model can outperform commercial models in specific scenarios. Additionally, we evaluate DriveMM against the open-source generalist model LLaVA-OneVision (LLaVA-OV) [19], which has a similar structure to ours and can process different types of data inputs. LLaVA-OV exhibits a certain ability to handle single-image/video data such as LingoQA, but its performance is poor in multi-view videos, such as NuInstruct. This limitation may be attributed to the fact that LLaVA-OV is not pre-trained on multi-view videos. To verify the effectiveness of our prompt and AD dataset, we equip LLaVA-OV with the perspective-aware prompt and fine-tune it on our enhanced AD datasets. The resulting model, named Drive-OV, exhibits substantial improvement across all benchmarks, demonstrating the compatibility of our prompt and dataset. Specifically, it achieves a performance similar to DriveMM on LingoQA. Drive-OV is also pre-trained on a significant amount of videos, while LingoQA primarily focuses on simple scene understanding and action description tasks without involving detection and prediction. Consequently, Drive-OV achieves comparable performance to DriveMM. However, on other more complex datasets, DriveMM outperforms Drive-OV, particularly in the datasets involving detection and grounding tasks, such as DriveLM and NuInstruct. This highlights our multi-capability pre-trained base model has greater potential to process complex AD scenes.

To verify the model’s generalization ability, we conduct zero-shot experiments on the popular BDD-X [17] dataset. The test set of BDD-X consists of 698 driving videos. Each video includes around 3-4 actions, all annotated with action descriptions and explanations. Following the previous works such as CODA-LM [22] and DriveLM [46], we utilize GPT-4o [16] as the evaluator to assesses the congruence between the model’s predictions and the ground truths. For comparison, we train the specialist models on each single dataset and prompt the model to predict the action description and explanation on the unseen dataset BDD-X. As shown in Tab. 4, our generalist DriveMM demonstrates better generalization performance than the specialist models. This validates that the general-purpose all-in-one DriveMM can better adapt to new driving scenarios and tasks. Please refer to the supplementary material for more details.

ID	Perspective-aware prompt	Question-answer augmentation	Question-answer standardization	Multi-capacity pre-training	CODA-LM	MAPLM	DriveLM	LingoQA	OmniDrive	NuInstruct
1	✗	✗	✗	✗	62.56	70.65	46.85	70.39	49.27	33.71
2	✓	✗	✗	✗	62.86	70.18	47.41	68.60	49.69	34.51
3	✓	✓	✗	✗	63.93	72.71	59.67	70.80	49.70	34.77
4	✓	✓	✓	✗	63.90	74.05	60.54	71.20	50.23	42.44
5	✓	✓	✓	✓	64.18	76.67	61.30	69.20	50.25	46.30

Table 5. Ablation on the proposed components. We report the average score for each autonomous driving dataset.

Fine-tuning Data	CODA-LM	MAPLM	DriveLM	LingoQA	OmniDrive	NuInstruct
Individual dataset	64.13	74.02	60.91	67.40	49.22	45.06
Mixed dataset	64.18	76.67	61.30	69.20	50.25	46.30

Table 6. Ablation on multi-dataset fine-tuning. Individual dataset and mixed dataset denote single dataset and a combination of six datasets.

### 4.3. Ablation Study

In this section, we conduct ablation studies to thoroughly validate the efficacy of the proposed components, including the perspective-aware prompt, Question-Answer (QA) enhancement, multi-capacity pre-training and multi-dataset fine-tuning. We report the average score for each dataset.

**Perspective-aware prompt.** We assess the effectiveness of the perspective-aware prompt in processing autonomous driving data. As evidenced by the comparison between Exp. 1 and 2 in Tab. 5, the perspective-aware prompt enhances the performances on the multi-view datasets such as DriveLM, OmniDrive and NuInstruct, which involve inputs with multiple perspectives and perspective-related questions. This validation confirms that the inclusion of the perspectives and data format in the instruction enables the model to better capture the perspective-related features, and thus enhances the model’s comprehension of multi-view data and spatial perception.

**Question-answer enhancement.** Here, we conduct an ablation study on both QA augmentation and QA standardization. The experimental results from Exp. 2 and 3 in Tab. 5 demonstrate significant improvements in datasets with constrained question styles, such as CODA-LM and MAPLM (+1.07 and +2.53), as well as with multiple-choice questions like DriveLM (+12.26) after applying QA augmentation. Furthermore, Exp. 3 and 4 in Tab. 5 reveal that QA standardization effectively enhances the performance in datasets like DriveLM and NuInstruct (+0.87 and +7.67), which involve detection tasks. Notably, the performances of all datasets exhibited improvement after applying QA augmentation and QA standardization. These results show that such two QA enhancement technologies effectively promote collaboration between multiple datasets.

**Multi-capacity pre-training.** To investigate the efficacy of the multi-type data and perception data in pre-training, we exclude the multi-capacity pre-training stage in Exp. 4 as shown in Tab. 5. By comparing the final results in Exp. 5, we observe a decline in performance in datasets involving video input and localization tasks, such as NuInstruct and DriveLM. This suggests that multi-capacity pre-training is vital for enhancing the model’s capacity to handle diverse visual information and enhance perception capabilities in complex autonomous driving scenes.

**Multi-dataset fine-tuning.** To validate the mutual enhancement between different datasets, we respectively train the model on the mixed autonomous driving dataset and the individual one. As illustrated in Tab. 6, the model trained on the mixed dataset outperforms those trained on the individual dataset. This observation underscores the efficacy of mixed training in harnessing the complementary information inherent in diverse datasets, thereby enhancing the model’s overall performance and robustness. Moreover, this approach mitigates the risk of overfitting any single dataset, promoting a more balanced learning process that captures a wider range of features and patterns.

### 5. Conclusion

In this paper, we present an all-in-one large multimodal autonomous driving model, DriveMM, which can handle various types of data and perform multiple driving tasks in real-world scenarios, demonstrating excellent generality and robustness. To our knowledge, we are the first to develop a comprehensive model for AD and evaluate the model across multiple datasets in various AD scenarios. By augmenting and standardizing several open-source datasets and designing data-related prompts, we conduct multi-step pre-training and fine-tuning of the model from scratch. DriveMM achieves state-of-the-art performance across various data and tasks in the real-world scenarios.

## References

- [1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 3
- [2] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 2023. 4
- [3] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *CVPR*, pages 11621–11631, 2020. 2, 4, 7, 12
- [4] Xu Cao, Tong Zhou, Yunsheng Ma, Wenqian Ye, Can Cui, Kun Tang, Zhipeng Cao, Kaizhao Liang, Ziran Wang, James M Rehg, et al. Maplm: A real-world large-scale vision-language benchmark for map and traffic scene understanding. In *CVPR*, pages 21819–21830, 2024. 1, 2, 3, 5, 6
- [5] Guiming Hardy Chen, Shunian Chen, Ruifei Zhang, Junying Chen, Xiangbo Wu, Zhiyi Zhang, Zhihong Chen, Jianquan Li, Xiang Wan, and Benyou Wang. Allava: Harnessing gpt4v-synthesized data for a lite vision-language model. *arXiv preprint arXiv:2402.11684*, 2024. 12
- [6] Li Chen, Penghao Wu, Kashyap Chitta, Bernhard Jaeger, Andreas Geiger, and Hongyang Li. End-to-end autonomous driving: Challenges and frontiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 1
- [7] Shimin Chen, Xiaohan Lan, Yitian Yuan, Zequn Jie, and Lin Ma. Timemarker: A versatile video-lm for long and short video understanding with superior temporal localization ability. *arXiv preprint arXiv:2411.18211*, 2024. 3
- [8] Xi Chen, Xiao Wang, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, et al. Pali: A jointly-scaled multilingual language-image model. *arXiv preprint arXiv:2209.06794*, 2022. 12
- [9] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *CVPR*, pages 24185–24198, 2024. 4
- [10] Mingfei Cheng, Yuan Zhou, and Xiaofei Xie. Behavexplor: Behavior diversity guided testing for autonomous driving systems. In *Proceedings of the 32nd ACM SIGSOFT International Symposium on Software Testing and Analysis*, pages 488–500, 2023. 1
- [11] Can Cui, Yunsheng Ma, Xu Cao, Wenqian Ye, and Ziran Wang. Drive as you speak: Enabling human-like interaction with large language models in autonomous vehicles. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 902–909, 2024. 1
- [12] Wenhao Ding, Chejian Xu, Mansur Arief, Haohong Lin, Bo Li, and Ding Zhao. A survey on safety-critical driving scenario generation—a methodological perspective. *IEEE Transactions on Intelligent Transportation Systems*, 24(7): 6971–6988, 2023. 1
- [13] Xinpeng Ding, Jianhua Han, Hang Xu, Xiaodan Liang, Wei Zhang, and Xiaomeng Li. Holistic autonomous driving understanding by bird's-eye-view injected multi-modal large models. In *CVPR*, pages 13668–13677, 2024. 1, 2, 3, 5, 6
- [14] Zhijian Huang, Sihao Lin, Guiyu Liu, Mukun Luo, Chaoqiang Ye, Hang Xu, Xiaojun Chang, and Xiaodan Liang. Fuller: Unified multi-modality multi-task 3d perception via multi-level gradient calibration. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3502–3511, 2023. 1
- [15] Zhijian Huang, Tao Tang, Shaoxiang Chen, Sihao Lin, Zequn Jie, Lin Ma, Guangrun Wang, and Xiaodan Liang. Making large language models better planners with reasoning-decision alignment. In *European Conference on Computer Vision*, pages 73–90. Springer, 2025. 1, 3
- [16] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024. 6, 7
- [17] Jinkyu Kim, Anna Rohrbach, Trevor Darrell, John Canny, and Zeynep Akata. Textual explanations for self-driving vehicles. In *ECCV*, pages 563–578, 2018. 1, 2, 7
- [18] Bo Li, Hao Zhang, Kaichen Zhang, Dong Guo, Yuanhan Zhang, Renrui Zhang, Feng Li, Ziwei Liu, and Chunyuan Li. Llava-next: What else influences visual instruction tuning beyond data, 2024. 4, 12
- [19] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024. 4, 6, 7
- [20] Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models. *arXiv preprint arXiv:2407.07895*, 2024. 12
- [21] Jiajhan Li and Tong Lu. Driving with internvl. 2024. 2, 6, 7
- [22] Yanze Li, Wenhua Zhang, Kai Chen, Yanxin Liu, Pengxiang Li, Ruiyuan Gao, Lanqing Hong, Meng Tian, Xinhai Zhao, Zhenguo Li, et al. Automated evaluation of large vision-language models on self-driving corner cases. *arXiv preprint arXiv:2404.10595*, 2024. 1, 2, 5, 6, 7
- [23] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, pages 740–755. Springer, 2014. 4, 12
- [24] Fanfan Liu, Feng Yan, Liming Zheng, Chengjian Feng, Yiyang Huang, and Lin Ma. Robouniview: Visual-language model with unified view representation for robotic manipulation. *arXiv preprint arXiv:2406.18977*, 2024. 3
- [25] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306, 2024. 12
- [26] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *NeurIPS*, 36, 2024. 1, 3, 4, 6, 12

- [27] Yongkang Liu and John HL Hansen. Towards complexity level classification of driving scenarios using environmental information. In *2019 IEEE Intelligent Transportation Systems Conference (ITSC)*, pages 810–815. IEEE, 2019. 1
- [28] Yinpeng Liu, Jiawei Liu, Xiang Shi, Qikai Cheng, Yong Huang, and Wei Lu. Let’s learn step by step: Enhancing in-context learning ability with curriculum learning. *arXiv preprint arXiv:2402.10738*, 2024. 2
- [29] Daqin Luo, Chengjian Feng, Yuxuan Nong, and Yiqing Shen. Autom3l: An automated multimodal machine learning framework with large language models. In *ACM MM*, pages 8586–8594, 2024. 3
- [30] Srikanth Malla, Chiho Choi, Ishi Dwivedi, Joon Hee Choi, and Jiachen Li. Drama: Joint risk localization and captioning in driving. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 1043–1052, 2023. 2
- [31] Jiageng Mao, Yuxi Qian, Junjie Ye, Hang Zhao, and Yue Wang. Gpt-driver: Learning to drive with gpt. *arXiv preprint arXiv:2310.01415*, 2023. 3
- [32] Jiageng Mao, Yuxi Qian, Junjie Ye, Hang Zhao, and Yue Wang. Gpt-driver: Learning to drive with gpt. *arXiv preprint arXiv:2310.01415*, 2023. 1
- [33] Jiageng Mao, Junjie Ye, Yuxi Qian, Marco Pavone, and Yue Wang. A language agent for autonomous driving. *arXiv preprint arXiv:2311.10813*, 2023. 1
- [34] Ana-Maria Marcu, Long Chen, Jan Hünermann, Alice Karnsund, Benoit Hanotte, Prajwal Chidananda, Saurabh Nair, Vijay Badrinarayanan, Alex Kendall, Jamie Shotton, et al. Lingoqa: Video question answering for autonomous driving. In *ECCV*, 2024. 1, 2, 5, 6
- [35] AI Meta. Introducing llama 3.1: Our most capable models to date. *Meta AI Blog*, 2024. 12
- [36] Jessica Hafetz Mirman. A dynamical systems perspective on driver behavior. *Transportation research part F: traffic psychology and behaviour*, 63:193–203, 2019. 1
- [37] Ming Nie, Renyuan Peng, Chunwei Wang, Xinyue Cai, Jianhua Han, Hang Xu, and Li Zhang. Reason2drive: Towards interpretable and chain-based reasoning for autonomous driving. In *ECCV*, pages 292–308. Springer, 2025. 3
- [38] Chenbin Pan, Burhaneddin Yaman, Tommaso Nesti, Abhirup Mallik, Alessandro G Allievi, Senem Velipasalar, and Liu Ren. Vlp: Vision language planning for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14760–14769, 2024. 1
- [39] SungYeon Park, MinJae Lee, JiHyuk Kang, Hahyeon Choi, Yoonah Park, Juhwan Cho, Adam Lee, and DongKyu Kim. Vlaad: Vision and language assistant for autonomous driving. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 980–987, 2024. 2
- [40] Tianwen Qian, Jingjing Chen, Linhai Zhuo, Yang Jiao, and Yu-Gang Jiang. Nuscenes-qa: A multi-modal visual question answering benchmark for autonomous driving scenario. In *AAAI*, pages 4542–4550, 2024. 2
- [41] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Worthzman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *NeurIPS*, 35:25278–25294, 2022. 4
- [42] Hao Shao, Yuxuan Hu, Letian Wang, Guanglu Song, Steven L Waslander, Yu Liu, and Hongsheng Li. Lmdrive: Closed-loop end-to-end driving with large language models. In *CVPR*, pages 15120–15130, 2024. 1, 3
- [43] Shuai Shao, Zeming Li, Tianyuan Zhang, Chao Peng, Gang Yu, Xiangyu Zhang, Jing Li, and Jian Sun. Objects365: A large-scale, high-quality dataset for object detection. In *ICCV*, pages 8430–8439, 2019. 4, 12
- [44] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, 2018. 4, 12
- [45] Chonghao Sima, Katrin Renz, Kashyap Chitta, Li Chen, Hanxue Zhang, Chengen Xie, Jens Beißenwenger, Ping Luo, Andreas Geiger, and Hongyang Li. Drivelm: Driving with graph visual question answering. In *ECCV*, 2024. 1, 2, 5, 6
- [46] Xiaoyu Tian, Junru Gu, Bailin Li, Yicheng Liu, Yang Wang, Zhiyong Zhao, Kun Zhan, Peng Jia, Xianpeng Lang, and Hang Zhao. Drivevlm: The convergence of autonomous driving and large vision-language models. *arXiv preprint arXiv:2402.12289*, 2024. 2, 7
- [47] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 4
- [48] Pengqin Wang, Meixin Zhu, Hongliang Lu, Hui Zhong, Xianda Chen, Shaojie Shen, Xuesong Wang, and Yinhai Wang. Bevgpt: Generative pre-trained large model for autonomous driving prediction, decision-making, and planning. *arXiv preprint arXiv:2310.10357*, 2023. 1
- [49] Shihao Wang, Zhiding Yu, Xiaohui Jiang, Shiyi Lan, Min Shi, Nadine Chang, Jan Kautz, Ying Li, and Jose M Alvarez. Omnidrive: A holistic llm-agent framework for autonomous driving with 3d perception, reasoning and planning. *arXiv preprint arXiv:2405.01533*, 2024. 1, 2, 3, 5, 6
- [50] Wenhai Wang, Jiangwei Xie, ChuanYang Hu, Haoming Zou, Jianan Fan, Wenwen Tong, Yang Wen, Silei Wu, Hanming Deng, Zhiqi Li, et al. Driveqlm: Aligning multi-modal large language models with behavioral planning states for autonomous driving. *arXiv preprint arXiv:2312.09245*, 2023. 1, 3
- [51] Licheng Wen, Daocheng Fu, Xin Li, Xinyu Cai, Tao Ma, Pinlong Cai, Min Dou, Botian Shi, Liang He, and Yu Qiao. Dilu: A knowledge-driven approach to autonomous driving with large language models. *arXiv preprint arXiv:2309.16292*, 2023. 1, 3
- [52] Yiran Xu, Xiaoyin Yang, Lihang Gong, Hsuan-Chu Lin, Tz-Ying Wu, Yunsheng Li, and Nuno Vasconcelos. Explainable object-induced action decision for autonomous vehicles. In *CVPR*, pages 9523–9532, 2020. 2

- [53] Zhenhua Xu, Yujia Zhang, Enze Xie, Zhen Zhao, Yong Guo, Kwan-Yee K Wong, Zhenguo Li, and Hengshuang Zhao. Drivegpt4: Interpretable end-to-end autonomous driving via large language model. *IEEE Robotics and Automation Letters*, 2024. [1](#), [2](#), [3](#)
- [54] Zhenjie Yang, Xiaosong Jia, Hongyang Li, and Junchi Yan. A survey of large language models for autonomous driving. *arXiv preprint arXiv:2311.01043*, 2023. [1](#)
- [55] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *ICCV*, pages 11975–11986, 2023. [4](#), [12](#)
- [56] Lu Zhang, KongJian Qin, BoYa Zhou, and HuaSen Wang. A method for evaluating the complexity of test scenarios for autonomous vehicles. In *Third International Conference on Mechanical, Electronics, and Electrical and Automation Control (METMS 2023)*, pages 291–301. SPIE, 2023. [1](#)
- [57] Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Video instruction tuning with synthetic data. *arXiv preprint arXiv:2410.02713*, 2024. [12](#)
- [58] Zeyu Zhang, Xiaohe Bo, Chen Ma, Rui Li, Xu Chen, Quanyu Dai, Jieming Zhu, Zhenhua Dong, and Ji-Rong Wen. A survey on the memory mechanism of large language model based agents. *arXiv preprint arXiv:2404.13501*, 2024. [1](#)
- [59] Yunsong Zhou, Linyan Huang, Qingwen Bu, Jia Zeng, Tianyu Li, Hang Qiu, Hongzi Zhu, Minyi Guo, Yu Qiao, and Hongyang Li. Embodied understanding of driving scenarios. *arXiv preprint arXiv:2403.04593*, 2024. [3](#)

# DriveMM: All-in-One Large Multimodal Model for Autonomous Driving

## Supplementary Material

### A. Implementation Details

In this section, we elaborate on the more detailed implementations for DriveMM and the experiments in Sec. 3.2.

**Model architecture.** We adapt SigLIP [55] as the vision encoder, which is pre-trained on WebLI [8] with a resolution of  $384 \times 384$ . We use a 2-layer MLP [25] as the projector to project the image features into the word embedding space. For the language model, we choose Llama-3.1 [35] 8B, which uses a tokenizer with a vocabulary of 128K tokens. Our model is trained on sequences of 8,192 tokens.

**Experiment setting.** Here we describe the training details of DriveMM. We adopt a curriculum learning approach to progressively train DriveMM as introduced in Sec. 3.4.

- *Stage-1: Language-image alignment.* We use LCS-558K [26] to align the visual patch features into the word embedding space. During this stage, we train only the projector while keeping the other components frozen. The learning rate is set to  $1 \times 10^{-3}$  and the training is conducted for 1 epoch with a batch size of 512.
- *Stage-2: Single-image pre-training.* In this stage, we use single-image data to improve model’s image comprehension capability. We utilize the recaptioned BLIP558K [26], COCO118K [18] and CC3M [44] datasets to improve the model. Meanwhile, we use language data Evo-Instruct [5] to balance the model’s language understanding ability. At this stage, the dataset comprises 3M single-image data and 143K language data. We fine-tune the entire model using a batch size of 256 for 1 epoch. The learning rate for the vision encoder is set to  $2 \times 10^{-6}$ , while the learning rate for both the projector and LLM is  $1 \times 10^{-5}$ .
- *Stage-3: Multi-capacity pre-training.* In this stage, our primary objective is to enhance the model’s reasoning and perception capabilities, and equip the model with the ability to handle diverse data formats. To achieve this, we use various multimodal data and perception data, including 1.5M single images, 760K multi-view images, 501K single videos, and 145K multi-view videos. Specifically, the single-image data consists of the multimodal data from [18] and the perception data from COCO [23] and Object365 [43]. The multi-view image data includes the multimodal data LLaVA-NeXT-Interleave [20] and the perception data nuScenes [3]. The single-view video data is derived from the works [20, 57]. Given the scarcity of multi-view videos in the available data, we generate

the multi-view perception data using nuScenes [3], with each view consisting of 5 frames. We fine-tune the entire model with a batch size of 256 for 1 epoch, maintaining the same learning rates as Stage-2.

- *Stage-4: Driving fine-tuning.* In the final stage, we employ a diverse array of high-quality driving datasets to fine-tune DriveMM for AD tasks. We compile six public AD datasets, which include single image (CODA-LM), multi-view images (MAPLM, DriveLM), single video (LingoQA), and multi-view videos (OmniDrive, NuInstruct), amounting to a total of 1.5M. Specifically, we augment the CODA-LM dataset, expanding it from 36,896 to 184,480 samples, and the MAPLM dataset from 47,485 to 94,970 samples. Additionally, we standardize the DriveLM and NuInstruct datasets to ensure uniformity across the data. In this stage, we use the same batch size and learning rate as Stage-2.

### B. Prompt Design

In the question-answer augmentation, we use the prompt presented in Tab. S1 for GPT-4o-mini to enhance and improve the question-answer pairs.

```
Messages = [
    {"role": "system", "content": f""" You are an English
improver. """},
    {"role": "user", "content": f""" I have a question and
its corresponding answer. I need your assistance in revis-
ing and refining them. Please make some changes to the
written content while preserving the meaning. The ques-
tion and answer that require modifications are: {QA}.
Please provide the revised question and answer in the for-
mat: Question: <question> Answer: <answer>. """}]
```

Table S1. The prompt for question-answer augmentation.

### C. Qualitative Results

Here we present the qualitative results for analyzing the distinctions among various models. In Fig. S1-S16, we provide the visualization examples from each dataset, illustrating the general performance on the six benchmarks and the generalization ability in BDD-X. The specialist model is trained on individual dataset. The all-in-one model DriveMM is able to accurately identify and analyze key objects related to driving in images and videos, thereby enhancing its ability to handle diverse autonomous driving data and tasks.

**CODA-LM**

**User**

<image>. There is an image of traffic captured from the front view of the ego vehicle. Focus on objects influencing the ego car's driving behavior: vehicles (cars, trucks, buses, etc.), vulnerable road users (pedestrians, cyclists, motorcyclists), traffic signs (no parking, warning, directional, etc.), traffic lights (red, green, yellow), traffic cones, barriers, miscellaneous(debris, dustbin, animals, etc.). You must not discuss any objects beyond the seven categories above. Please describe each object's appearance, position, direction, and explain why it affects the ego car's behavior.

**DriveMM**

In the traffic image, there is a **black SUV** directly in front of the ego car, occupying the same lane and moving in the same direction. The presence of this SUV is significant as it dictates the following distance and speed that the ego car must maintain to ensure safe driving conditions. Additionally, there are **pedestrians on the sidewalk** to the right of the ego car, with one individual wearing a bright orange outfit and standing close to the curb. The proximity of these pedestrians to the road means that the ego car must remain vigilant and be prepared to stop or slow down if any of them unexpectedly enter the roadway. On the right side of the road, there is a **'No Entry' sign** for vehicles, accompanied by a blue sign indicating a detour or alternative route. This sign is crucial as it informs the ego car that it cannot proceed straight ahead in its current lane and must instead follow the detour to continue its journey. Furthermore, there are **orange traffic cones** placed on the right side of the road, partially obstructing the sidewalk and the edge of the road. These cones suggest that there may be road work or an obstruction ahead, which could affect the available driving space and necessitate caution from the ego car. There are no traffic lights, barriers, or other objects reported in the image that affect the driving behavior of the ego car.

**Specialist**

In the traffic image, there is a **dark-colored SUV** directly in front of the ego car, occupying the same lane and moving in the same direction. The presence of this vehicle requires the ego car to maintain a safe following distance and be prepared to adjust speed or change lanes if necessary. On the right side of the road, there are **pedestrians** walking parallel to the road. The ego car must be cautious and ready to stop or slow down in case any pedestrian steps onto the road. Additionally, there is a **cyclist** on the right side of the road, moving in the same direction as the ego car. The ego car should give the cyclist enough space and be prepared to slow down or change lanes if the cyclist's path intersects with the car's path. There is also a **'No Entry' sign** on the right side of the road, indicating that the road ahead is closed to traffic. The ego car must not proceed forward and should look for an alternative route. Furthermore, there are traffic cones placed on the right side of the road, indicating a work zone or a closed area. The ego car should be cautious and may need to merge left if the right lane is obstructed. There are no traffic lights, barriers, or other objects present in the image that affect the driving behavior of the ego car.

**GPT-4o**

1. **Vehicle (Car)** - **Appearance**: **Dark-colored SUV**. - **Position**: Directly ahead of the ego car in the same lane. - **Direction**: Moving forward. - **Influence**: The ego car must maintain a safe following distance and adjust its speed according to the SUV's speed.
2. **Vulnerable Road Users (Pedestrians)** - **Appearance**: Several pedestrians, some walking and some standing. - **Position**: On the left side of the road, near the sidewalk, and one pedestrian on the right side near the 'Do Not Enter' sign.
3. **Traffic Sign (Do Not Enter)** - **Appearance**: Red circle with a white horizontal line and the text 'DO NOT ENTER.' - **Position**: On the right side of the road. - **Direction**: Facing the ego car. - **Influence**: The ego car must not enter the road indicated by the sign and should continue straight or follow any other permissible route.
4. **Miscellaneous (Debris)** - **Appearance**: Small debris scattered on the road. - **Position**: On the road surface, near the center. - **Direction**: Stationary. - **Influence**: The ego car should avoid the debris to prevent potential damage to the vehicle and ensure a smooth drive.

Figure S1. Visualization of CODA-LM. Key information is highlighted in **green**, while errors are marked in **red**.

CODA-LM



**User** <image>. There is an image of traffic captured from the front view of the ego vehicle. Please describe the object inside the red rectangle in the image and explain why it affect ego car driving.

**DriveMM** This object is a [traffic sign displaying a red circle with a white bar](#), indicating a prohibition of entry for vehicles in the direction the sign is facing. The ego car must not proceed forward or turn in the direction indicated by this sign to comply with traffic rules.

**Specialist** This is a traffic sign, specifically indicating [a no parking or stopping zone](#). It is important for the ego car to recognize this sign to avoid stopping or parking in this area, which could result in a traffic violation.

**GPT-4o** The object inside the red rectangle is [a pedestrian](#). The presence of the pedestrian affects the ego car's driving because the car must slow down or stop to ensure the safety of the pedestrian, adhering to traffic laws and preventing potential accidents.

Figure S2. Visualization of CODA-LM.

CODA-LM



**User**

<image>. There is an image of traffic captured from the front view of the ego vehicle. Focus on objects influencing the ego car's driving behavior: vehicles (cars, trucks, buses, etc.), vulnerable road users (pedestrians, cyclists, motorcyclists), traffic signs (no parking, warning, directional, etc.), traffic lights (red, green, yellow), traffic cones, barriers, miscellaneous(debris, dustbin, animals, etc.). You must not discuss any objects beyond the seven categories above. Please provide driving suggestions for the ego car based on the current scene.

**DriveMM**

The ego car should maintain a safe following distance from the **vehicle ahead**, adhere to the **20 km/h speed limit**, and be prepared to slow down or stop for the **road workers**. It should also be ready to merge left if necessary due to the **traffic cones** indicating a potential lane closure.

**Specialist**

The ego car should maintain a safe following distance from the vehicles ahead, be prepared to slow down or stop for the road workers, and adhere to the **20 km/h speed limit**. It should also be ready to change lanes if necessary, while being cautious of the **cyclist** and giving them ample space.

**GPT-4o**

Based on the image, here are the driving suggestions for the ego car: 1. **Vehicles**: There is a **car ahead** in the same lane. Maintain a safe following distance and be prepared to adjust speed based on the car's movements. 2. **Vulnerable Road Users**: There are **pedestrians** on the right side of the road. Drive cautiously and be prepared to stop if any pedestrian steps onto the road. 3. **Traffic Signs**: There are **no visible traffic signs** directly influencing the ego car's behavior. 4. **Traffic Lights**: There are no traffic lights visible in the image. 5. **Traffic Cones**: There are **traffic cones** on the right side of the road, indicating a potential work zone or hazard. Slow down and proceed with caution. 6. **Barriers**: There are no barriers directly influencing the ego car's behavior. 7. **Miscellaneous**: There are no miscellaneous objects directly influencing the ego car's behavior. **Driving Suggestions**: - Maintain a safe following distance from the car ahead. - Slow down and be prepared to stop for pedestrians. - Proceed with caution near the traffic cones, as they may indicate a work zone or hazard.

Figure S3. Visualization of CODA-LM.

**MAPLM**

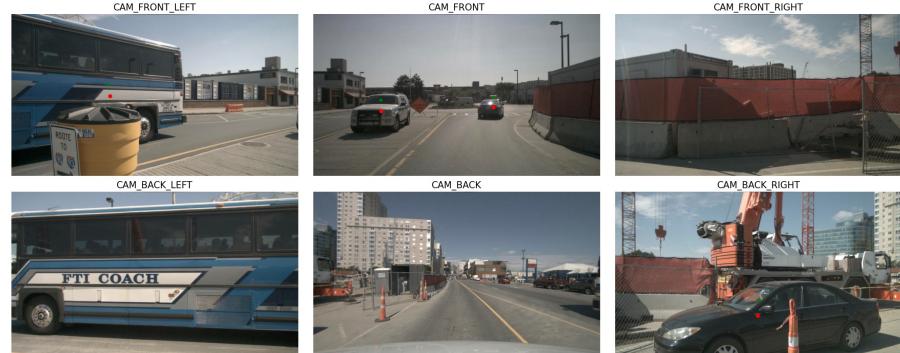
**User** 1: <image> 2: <image> 3: <image> 4: <image>. These four images are the point cloud bev, camera front view, camera back left view, camera right view of the ego vehicle. How many lanes on the current road?

**DriveMM** 3.

**Specialist** 4.

**GPT-4o** The current road has **two lanes**.

Figure S4. Visualization of MAPLM.

**DriveLM**

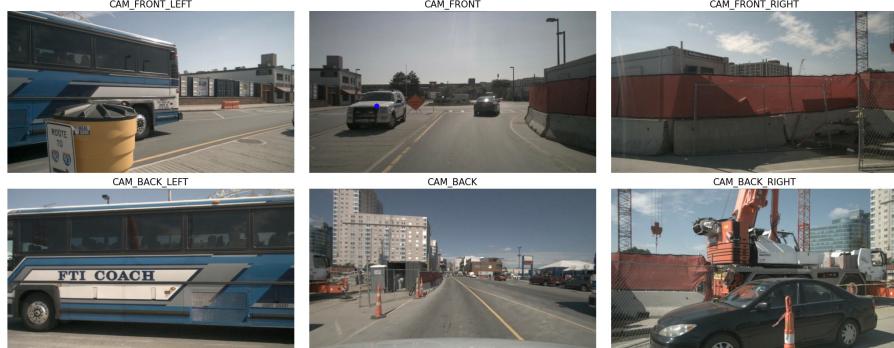
**User** 1: <image> 2: <image> 3: <image> 4: <image> 5: <image> 6: <image>. These six images are the front view, front left view, front right view, back view, back left view and back right view of the ego vehicle. What are the important objects in the current scene? Those objects will be considered for the future reasoning and driving decision.

**DriveMM** There is a black car to the back right of the ego vehicle, a black sedan to the front of the ego vehicle, a white SUV to the front of the ego vehicle, and a blue and white bus to the front left of the ego vehicle. The IDs of these objects are <c1,CAM\_BACK\_RIGHT,800.8,599.288>, <c2,CAM\_FRONT,1004.192,463.736>, <c3,CAM\_FRONT,375.008,479.288>, and <c4,CAM\_FRONT\_LEFT,535.792,393.336>.

**Specialist** There is a black car to the back right of the ego vehicle, a black sedan to the front of the ego vehicle, a white SUV to the front of the ego vehicle, and a blue and white bus to the front left of the ego vehicle. The IDs of these objects are <c1,CAM\_BACK\_RIGHT,794.2,687.5>, <c2,CAM\_FRONT,1004.2,517.5>, <c3,CAM\_FRONT,377.5,539.2>, and <c4,CAM\_FRONT\_LEFT,548.3,455.0>.

**GPT-4o** The important objects in the current scene that should be considered for future reasoning and driving decisions are: 1. Vehicles (including the **white car**, **black car**, and **buses**) 2. Construction barriers and cones 3. Construction equipment (e.g., crane) 4. Road signs and traffic signs 5. Buildings and structures along the road These objects are crucial for understanding the driving environment and making safe driving decisions.

Figure S5. Visualization of DriveLM.

**DriveLM****User**

1: <image> 2: <image> 3: <image> 4: <image> 5: <image> 6: <image>. These six images are the front view, front left view, front right view, back view, back left view and back right view of the ego vehicle. Would <c3,CAM\_FRONT,373.7,526.2> be in the moving direction of the ego vehicle?

**DriveMM** No.**Specialist** No.**GPT-4o**

Yes, <c3,CAM\_FRONT,373.7,526.2> would be in the moving direction of the ego vehicle. The CAM\_FRONT camera captures the front view, which aligns with the direction the vehicle is moving.

Figure S6. Visualization of DriveLM.

**LingoQA****User**

<image>. There is a video of traffic captured from the front view of the ego vehicle. Is there a traffic light in the vicinity? If so, what color is it displaying?

**DriveMM** Yes, there is a traffic light in the vicinity and it is displaying red.**Specialist** Yes, there is a traffic light and it is currently displaying red.**GPT-4o** Yes, there is a traffic light in the vicinity. It is displaying red.

Figure S7. Visualization of LingoQA.

LingoQA



**User** <image>. There is a video of traffic captured from the front view of the ego vehicle. What is the current action and its justification? Answer in the form \"action, justification\".

**DriveMM** I am **maintaining speed** and steering right to keep a safe distance to the adjacent lane on the left.

**Specialist** I am **accelerating** because there is a gap in the traffic behind the car ahead, and it is safe to do so.

**GPT-4o** Driving, clear road ahead.

Figure S8. Visualization of LingoQA.

OmniDrive



**User** 1: <image> 2: <image> 3: <image> 4: <image> 5: <image> 6: <image>. These six videos are the front view, front left view, front right view, back view, back left view and back right view of the ego vehicle. Is there any indication of a special event happening near the dome-like structure?

**DriveMM** There is no clear indication of a special event happening near the dome-like structure, but the presence of **tents** and a parked truck suggests some activity.

**Specialist** Here is no clear indication of a special event happening near the dome-like structure, but the presence of **tents** and **parked vehicles** suggests some sort of activity or event

**GPT-4o** Yes, there is an indication of a special event happening near the dome-like structure. The presence of **large tents** and **temporary structures** suggests that an event might be taking place.

Figure S9. Visualization of OmniDrive.

### OmniDrive



**User**

1: <image> 2: <image> 3: <image> 4: <image> 5: <image> 6: <image>. These six videos are the front view, front left view, front right view, back view, back left view and back right view of the ego vehicle. Given the proximity of other vehicles, what would be a safe driving practice right now?

**DriveMM**

A safe driving practice would be to **remain stationary**, maintain a safe distance from the vehicle ahead, and be prepared to proceed when the traffic light changes.

**Specialist**

A safe driving practice would be to maintain a safe following distance from the vehicle ahead, **stay alert** for any sudden stops or changes in traffic flow, and be prepared to react accordingly.

**GPT-4o**

Given the proximity of other vehicles, a safe driving practice right now would be to **remain stationary** and wait for a clear and safe opportunity to move. This will help avoid potential collisions with the nearby vehicles.

Figure S10. Visualization of OmniDrive.

### NuInstruct



**User**

1: <image> 2: <image> 3: <image> 4: <image> 5: <image> 6: <image>. These six videos are the front view, front left view, front right view, back view, back left view and back right view of the ego vehicle. What is the closest object in the right of the ego car? If any, please list all objects as format <class>[c,x1,y1,x2,y2], c is the class label, c indicates the camera view, and x1, y1, x2, y2 are the coordinates of the top-left and bottom-right corners of the bounding box.

**DriveMM**

<car>[CAM\_FRONT\_RIGHT,33.000,51.000,66.000,80.000];

**Specialist**

<pedestrian>[CAM\_FRONT\_RIGHT,84.000,45.000,91.000,68.000];

**GPT-4o**

<Car><3, 264, 144, 432, 288>

Figure S11. Visualization of NuInstruct.

**NuInstruct****User**

1: <image> 2: <image> 3: <image> 4: <image> 5: <image> 6: <image>. These six videos are the front view, front left view, front right view, back view, back left view and back right view of the ego vehicle. How many <vehicle>s in the back of the ego car?

**DriveMM**

1

**Specialist**

2

**GPT-4o**

There is one vehicle in the back of the ego car.

Figure S12. Visualization of NuInstruct.

**NuInstruct****User**

1: <image> 2: <image> 3: <image> 4: <image> 5: <image> 6: <image>. These six videos are the front view, front left view, front right view, back view, back left view and back right view of the ego vehicle. Does any objects approach the ego car? If any, please list all objects as format <class>[c,x1,y1,x2,y2], class is the class label, c indicates the camera view, and x1, y1, x2, y2 are the coordinates of the top-left and bottom-right corners of the bounding box.

**DriveMM**

<truck>[CAM\_BACK\_RIGHT,56.250,0.000,100.000,100.000];

**Specialist**

There are no objects that would have risks for the ego car.

**GPT-4o**

1. Person[3, 200, 150, 300, 400] 2. Person[3, 300, 150, 400, 400]

Figure S13. Visualization of NuInstruct.

### NuInstruct



#### User

1: <image> 2: <image> 3: <image> 4: <image> 5: <image> 6: <image>. These six videos are the front view, front left view, front right view, back view, back left view and back right view of the ego vehicle. What is the next motion for the ego? And give the reason. If any, please list all objects as format <class>[c,x1,y1,x2,y2], class is the class label, c indicates the camera view, and x1, y1, x2, y2 are the coordinates of the top-left and bottom-right corners of the bounding box.

#### DriveMM

There are <car>[CAM\_FRONT,70.812,51.000,100.000,100.000];<car>[CAM\_FRONT\_RIGHT,0.000,50.556,12.750,100.000]; on\_coming the ego car. However, they all keep a safe distance from ego vehicles. Hence, The ego would keep speed and go straightforward.

#### Specialist

There are no objects that would have risks for the ego car. Hence, The ego would keep speed and go straightforward.

#### GPT-4o

The next motion for the ego vehicle should be to proceed straight. Reason: The front view image shows a green traffic light, indicating that it is safe and legal for the ego vehicle to move forward.

Figure S14. Visualization of NuInstruct.

BDD-X



User

<image>. There is a video of traffic captured from the front view of the ego vehicle. Describe the current action of the ego car, and explain the cause of this car's action.

DriveMM

The ego car is **accelerating** because there is **a gap in the traffic behind the car ahead**.

Specialist

**CODA-LM**: The ego car is positioned at a **red traffic light**, which requires it to stop and wait for the signal to change to green before it can proceed.

**MAPLM**: There are **3 lanes in the image**, lane attributes from left to right are: motorway|motorway|bicycle lane.

**DriveLM**: The current action of the ego car is to **keep going at the same speed**. The cause of this car's action is to follow the traffic rules.

**LingoQA**: The current action of the ego car is **following the lane** and maintaining a safe distance from the adjacent lane on the right.

**Omnidrive**: The vehicle is currently proceeding straight ahead at a moderate speed, as the **traffic light is green** and the road ahead is clear.

**NuInstruct**: the ego no turn and **keep speed**

Figure S15. Visualization of BDD-X.

BDD-X



User

<image>. There is a video of traffic captured from the front view of the ego vehicle. Describe the current action of the ego car, and explain the cause of this car's action.

DriveMM

The ego car is currently **stationary** because the **traffic light is red**.

Specialist

**CODA-LM**: The ego car is halted at a **red traffic light**, which is clearly visible in the upper left corner of the image. The red light indicates that vehicles must stop and wait until the signal changes to green before proceeding. The ego car is positioned at an intersection, and the presence of other vehicles and a pedestrian in the vicinity suggests that it is a busy area. The ego car must remain stationary to comply with traffic regulations and to ensure the safety of pedestrians and other road users.

**MAPLM**: There are **3 lanes** in the image, lane attributes from left to right are: motorway|motorway|bicycle lane.

**DriveLM**: The action is **to remain stationary**, the cause is to follow the traffic rules.

**LingoQA**: The current action of the ego car is **accelerating** because the traffic light is green and the zebra crossing is clear.

**Omnidrive**: The vehicle is currently stopped at an intersection, waiting for the traffic light to turn green. This is because the **traffic light is red**, indicating that vehicles must halt until it changes to green.

**NuInstruct**: the ego no turn and **keep speed**

Figure S16. Visualization of BDD-X.