MAPLM: A Real-World Large-Scale Vision-Language Benchmark for Map and Traffic Scene Understanding

Xu Cao[1,2*], Tong Zhou[1*], Yunsheng Ma[3*], Wenqian Ye[4], Can Cui[3], Kun Tang[1], Zhipeng Cao[1]
Kaizhao Liang[5], Ziran Wang[3], James M. Rehg[2], Chao Zheng[1†]
[1]Tencent T Lab [2]University of Illinois Urbana-Champaign [3]Purdue University
[4]University of Virginia [5]SambaNova Systems, Inc.
xucao2@illinois.edu, chrisczheng@tencent.com

Abstract

Vision-language generative AI has demonstrated remarkable promise for empowering cross-modal scene understanding of autonomous driving and high-definition (HD) map systems. However, current benchmark datasets lack multi-modal point cloud, image, and language data pairs. Recent approaches utilize visual instruction learning and cross-modal prompt engineering to expand vision-language models into this domain. In this paper, we propose a new vision-language benchmark that can be used to finetune traffic and HD map domain-specific foundation models. Specifically, we annotate and leverage large-scale, broad-coverage traffic and map data extracted from huge HD map annotations, and use CLIP and LLaMA-2 / Vicuna to finetune a baseline model with instruction-following data. Our experimental results across various algorithms reveal that while visual instruction-tuning large language models (LLMs) can effectively learn meaningful representations from MAPLM-QA, there remains significant room for further advancements. To facilitate applying LLMs and multi-modal data into self-driving research, we will release our visual-language QA data, and the baseline models at GitHub.com/LLVM-AD/MAPLM.

1. Introduction

Recent breakthroughs in large language models (LLMs), with their incredible ability to reason [63] and interact with various tools [48], promise to bring a significant shift in the landscape of human-agent interaction [55, 65, 74]. They have also led to growing interest in multi-modal vision-language models (VLMs) [58, 78], which integrate and enhance the reasoning capabilities of LLMs with images, 3D LiDAR point clouds, videos, and audio and perform various

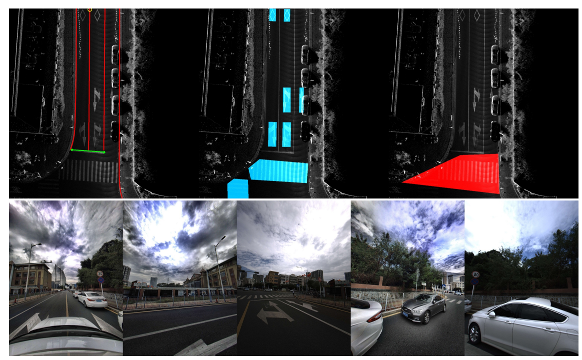


Figure 1. Panoramic 2D images, 3D LiDAR point cloud, and HD map annotations in MAPLM.

tasks such as image captioning, visual question answering (VQA), scene understanding. Besides, VLMs are used to align and map language with visual content, allowing language to play an important role in analyzing other signals and passing information to downstream LLMs [33].

In the autonomous driving (AD) industry, VLMs and LLMs have the potential to understand traffic scenes, thus enhancing the driving decision-making process and human-AI interaction of AD systems [10, 11, 22, 26, 76, 86]. By training on vast amounts of traffic scene data, they can glean insights from complex multi-modal driving resources such as map data, traffic laws, and incident reports [9]. This allows them to refine a vehicle's navigation and planning with safety and efficiency parameters, adapting to dynamic road conditions with an understanding that closely mirrors human intuition [4, 61].

However, while successful in the general domains, the current version of VLMs is less effective for traffic and driving scenarios as traffic data-text pairs contain diverse modalities across 3D LiDAR point clouds, panoramic 2D images, information from high-definition (HD) maps, are drastically different from the contexts and question-answer

*Equal Contribution
†Corresponding Author

Dataset	Year	QA	Caption	Scenario	Text	Modality			
						Context	Image	Point Cloud	HD Map Info
BDD-X [28]	2018	✗	✓	7K	26K	✓	✓	✗	✗
Talk2Car [13]	2019	✗	✓	34K	12K	✓	✓	✗	✗
SUTD-TrafficQA [71]	2021	✓	✗	10K	63K	✓	✓	✗	✗
DRAMA [39]	2023	✗	✓	18K	103K	✓	✓	✗	✗
nuScenes-QA [47]	2023	✓	✗	34K	460K	✓	✓	✓	✗
NuPrompt [68]	2023	✗	✓	34K	35K	✓	✓	✓	✗
DriveLM [8, 54]	2023	✓	✓	34K	375K	✓	✓	✗	✗
LINGO-QA [43]	2023	✓	✓	28K	420K	✓	✓	✗	✗
Rank2Tell [50]	2024	✗	✓	118	>118	✓	✓	✓	✗
NuScenes-MQA [27]	2024	✓	✓	34K	1.5M	✓	✓	✗	✗
MAPLM	2024	✗	✓	2M	2M	✓	✓	✓	✓
MAPLM-QA	2024	✓	✓	14K	61K	✓	✓	✓	✓

Table 1. Related datasets can be split into two types: (1) Add additional text annotations into existing datasets like nuScenes [6] (note with orange); (2) Collect independent data (note with blue).

pairs in the general domain. As a result, general-domain visual assistants may behave like laypersons, who would refrain from answering in-detailed traffic and map-related questions, or worse, produce incorrect responses or complete hallucinations in counting and localization questions [83]. Much progress has been made in traffic scene VQA and image captioning, but prior methods typically formulate the problem as a short information extraction task from single modality visual scenes and are secondary annotated from previous segmentation and object detection datasets [13]. Consequently, although LLMs and VLMs have demonstrated great potential for self-driving, map understanding applications [10, 34, 65, 76, 86], current research is often limited by data scale and ignores multi-modal alignments across different types of traffic scene data.

In this paper, we introduce MAPLM, a new benchmark to extend 3D LiDAR point clouds, panoramic 2D images, and HD map information into LLMs. The dataset contains a benchmark MAPLM-QA with 13,775 frames including image-text pairs extracted and annotated from HD maps. The scene in our dataset covers diverse image captioning and question-answer types. Inspired by recent work in instruction-tuning [33] and GPT-4V [44], we design a multi-modal baseline model for MAPLM.

The contributions of our work are the following:

- We propose MAPLM, a dataset consisting of millions of complex driving scenes and corresponding HD map text descriptions, and MAPLM-QA benchmark consisting of 14K frames containing multiple question-answer pairs for visual instruction tuning.
- To facilitate VLMs for driving and HD map scene understanding, we propose a novel multi-modal instruction tuning baseline model in the context of HD map information extraction for the MAPLM-QA benchmark.
- The baseline model of our MAPLM benchmark demonstrates superior traffic scene and map understanding performance compared to the state-of-the-art methods.

2. Related Works

2.1. Vision-Language Models

Researchers in computer vision have been actively exploring the use of VLMs for solving multi-modal tasks [31, 49, 82]. With the blooming of LLMs, one of the solutions is tool learning with foundation models [48]. By using tool learning, the LLMs can understand the user's intention and call related APIs like code generation to read data from different modalities when receiving the user's instruction, then generate responses by incorporating the results obtained from these APIs [53, 67, 77]. Another solution is finetuning or instruction tuning of fundamental large-scale VLMs [23, 46, 84] such as Flamingo [1] and MiniGPT4 [87]. Recent work LLaVA [33], Otter [30], InstructBLIP [12] develop instruction-following LLMs using the image-instruction tuning dataset, which proved the superiority of instruction tuning in multi-modal vision language tasks. However, current VLMs struggle to adapt to high-resolution and visually crowded images due to their absence of a visual search mechanism [69] and the limited visual grounding capabilities of CLIP [58].

2.2. Vision-Language Datasets for Driving Scenes

Since the task of visual question answering (VQA) was first proposed by [2], there have been plenty of VQA datasets for different research areas [5, 24, 36, 79, 80]. However, only a few of the VQA datasets focus on traffic scenes and HD map data which plays an important role in autonomous driving, and most of them lack key edge cases such as different weathers and locations [42]. In several pioneering datasets and benchmark papers, the authors have explored language-guided visual understanding tasks in driving scenes. These datasets can be split to two types: (1) Added additional texts for existing NuScenes [6] dataset such as Talk2Car [13], NuScenes-QA [47], NuScenes-MQA [27], DriveLM [8], and NuPrompt [68], NuInstruct [16]; (2) Independent collected datasets such as Rank2Tell [50], BDD-

X [28], SUTD-TrafficQA [71], DRAMA [39], and LINGO-QA [43]. However, limited by data scale and data quality, current datasets can not serve as useful benchmarks to evaluate multi-modal LLMs for driving scenes. Besides, the newest techniques like GPT-4V [44] in the general domain has already been trained with plenty of open-source traffic and driving scene datasets. Those vision language datasets annotated on nuScenes [6] can not serve as reliable benchmarks to validate existing models. Thus, we need new out of domain large-scale datasets and benchmarks that contain more corner cases of various traffic and driving scenarios and related HD map annotations.

2.3. LLMs for Autonomous Driving

LLMs have shown remarkable potential in complicated scenarios such as driving scene understanding and decision-making [10, 26, 38, 41]. Recent advancements focus on building visual-language models to generate driving policies such as DiLu [64], DriveGPT4 [72], GPT-Driver [40], HiLM-D [15], DriveMLM [60], and DriveVLM [57]. Talk2BEV [14] and LiDAR-LLM [75] also explored the connection between LLMs, VLMs and bird's-eye view (BEV), LiDAR point cloud in autonomous driving contexts. Besides, LLMs can also enhance the interaction between passengers and vehicles, improving the personalization and responsiveness of autonomous driving experiences [9, 20]. An equally crucial area of research is the development of language-guided closed-loop autonomous driving systems. These systems leverage multi-modal sensor data from simulators, as demonstrated by LimSim++ [19] and LMDrive [52]. Additionally, RAG-Driver [81] introduces a novel retrieval-augmented in-context learning approach, significantly enhancing the zero-shot generalization capabilities of driving LLMs. From the industry, Wayve proposed the first open-loop driving commentator LINGO-1 [62].

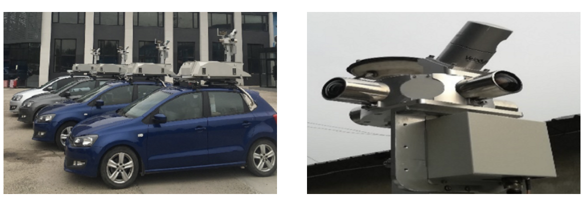


Figure 2. Device and data collection of MAPLM. We use collection cars to collect panoramic images and 3D LiDAR point clouds for the MAPLM benchmark.

3. Dataset: MAPLM

As we mentioned in the related work section, existing traffic and driving-related question-answering benchmarks are often limited by re-labeling previous publicly available datasets like NuScenes [6] or generated from simulators [17] and are hard to enable safe and detailed analysis required for real-world traffic scenes because their data contains few edge cases. To address this issue, we propose **MAPLM**, a dataset comprising real traffic scene data and related HD map context annotation. In addition to the visual data, we also released the MAPLM-QA benchmark, which consists of commonly used scene understanding questions across projected BEV images from 3D LiDAR point clouds, and panoramic 2D images.

3.1. Dataset Collection

We collect the MAPLM using HD map production automated vehicles including 6 cameras, a LiDAR scanner, installed at the tail at a 45-degree angle, focusing on scanning the road surface, and GPS/IMU integration systems (Figure 2) [56, 85]. The detailed collection parameters will be released in the MAPLM Dataset document. The raw 3D point cloud of MAPLM has the characteristics of high density, the apparent distinction between light and dark reflection intensity, and the apparent visual features of ground elements. MAPLM was collected from a variety of traffic scenarios, including highways, expressways, city roads, and rural roads, along with detailed intersection scenes, which ensure the MAPLM dataset contains enough driving edge cases [56].

3.2. Dataset Annotation

We split the annotation of MAPLM into two phases. In the first phase, we used our active learning-based multi-modal vision models for pre-labeling 3D LiDAR point clouds and panoramic RGB images, and then pre-labeling annotations were verified by a hired HD map annotation team. The production pipeline is similar to the traditional HD annotation process [18, 45, 51, 66]. We select the most representative scenarios among 3D LiDAR point clouds and panoramic images resulting in a total of 2 million frames of LiDAR point clouds and panoramic images (6 images). For each data point, we first extract text information from pre-labeled traffic scene annotations, including lane marking, ground marking, GPS, and road surface situation. Using HD map data, we also generate a list of text descriptions including (1) lane marking information in front of the car; (2) lane marking information behind the car; (3) stop line information around the car; (4) road sign information in front of the car; (5) road sign information behind the car; (6) cross zone around the car; (7) intersection zone around the car; (8) lane change zone around the car.

In the second phase, we hired another annotation team to verify the data caption annotation and create 13,775 new question-answer annotations from MAPLM as MAPLM-QA. Question-answer pairs target various tag dimensions, such as scene type, number and attributes of lanes, presence

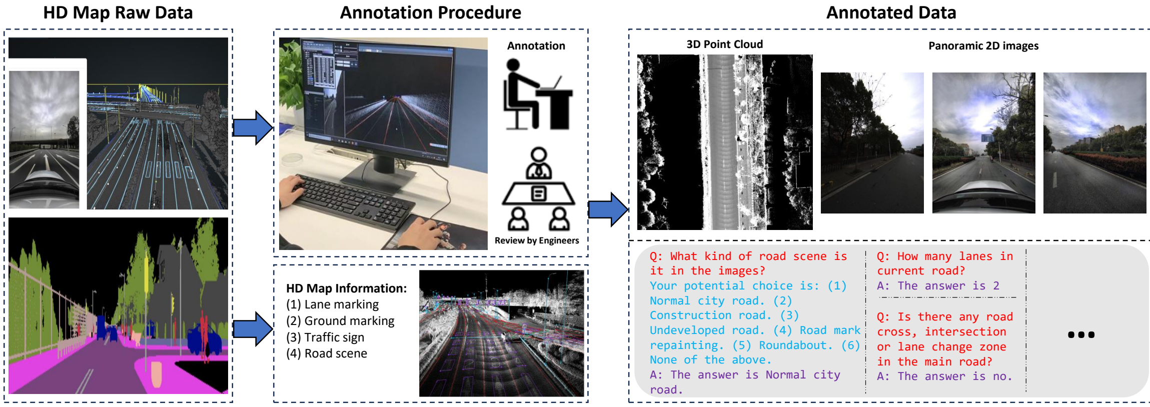


Figure 3. MAPLM and MAPLM-QA Dataset annotation procedure.

of intersections, etc. Sample questions are as follows (To simplify understanding, we employ abbreviations for each question type: **SCN** for road scene understanding; **QLT** for quality analysis of point cloud; **INT** for road intersection recognition; **LAN** for lane counting; **DES** for road and lane description. The number of questions is shown in brackets):

- **SCN**: What kind of road scene is it in the images? (13,775)
- **QLT**: What is the point cloud data quality in the current road area of this image? (13,775)
- **LAN**: How many lanes are on the current road? (13,775)
- **INT**: Is there any road cross, intersection, or lane change zone in the main road? (13,775)
- **DES**: Describe the lane attribute in the current road. (5,643)

The answers of **SCN**, **QLT**, and **INT** are from a set of choices, while the answers of **DES** are followed by **LAN**. It is used to describe the lane attribute in the current road scene, so the description will include two parts, (1) number of lanes; and (2) attribute of the lane. For example, Figure 4 shows the DES ground truth of a group of scenes.

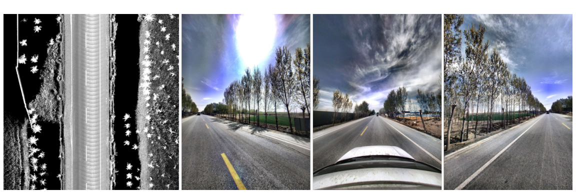


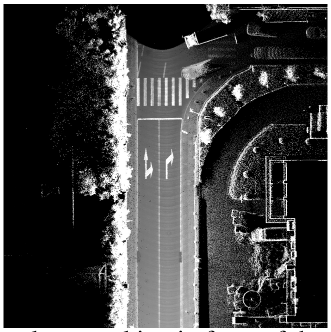
Figure 4. The DES of this scene is "There are 4 lanes in this scene, lane attributes from left to right are: bike lane | motorway lane | motorway lane | bike lane."

3.3. Data Statistics & Analysis

As Figure 6 describes, after removing general conversation words, the raw dataset contains well-balanced traffic and driving-related words. In Table 1, we compare our dataset MAPLM with other publicly available traffic, map, driving scene image captioning, and QA datasets. Below we will make a detailed comparison and explain the advantages of MAPLM from three aspects: scale, modality, and data quality. For data scales, MAPLM contains more scenarios than nuScenes-based datasets. Besides, MAPLM does not only include panoramic 2D images and projected BEV images from 3D LiDAR point clouds but also contains additional HD map information annotation which will be used as image captioning pretraining tasks for the CLIP visual encoder. The main edge cases in MAPLM are about geographical locations and lane attribute diversity based on HD map annotations (Table 2). The weather edge cases are also considered in data collection but the weather data statistics are not included in HD map annotations temporally. From our experiment, GPT-4V with zero-shot or few-shot inference can not perform well in MAPLM-QA, but a recently published tech report showed it can achieve good performance in traffic scenes in NuScenes [65]. A possible explanation is there is not enough out-of-domain knowledge included during the GPT-4V training.

Scene	Proportion
Highway Normal Road (city, rural area)	60% 40%
City Small Road / Alley Mountain Road Toll gate Tunnel	3.8% 4.7% 2.8% 6.6%
Road Construction	1.75%
Low Quality Data (lane marking occlusion, overlap, damage)	7.12%
Intersection	17.3%

Table 2. Geographical locations and lane attributes diversity in MAPLM.



Caption 1: One lane marking in front of the vehicle. From left to right: broken line.
Caption 2: Three lane marking behind the vehicle. From left to right: solid line, solid line, solid line.
Caption 3: One stop line around the vehicle.
Caption 4: Four road sign in front of the vehicle. From left to right, they are pavement arrow with go straight, turn right; pavement arrow with turn right; pedestrian crossing; pedestrian crossing.
Caption 5: No road sign behind the vehicle.
Caption 6: No crossroad or T-junction around the vehicle.
Caption 7: No small intersection zone around the vehicle.
Caption 8: No lane change zone around the vehicle.

Figure 5. Image caption description in MAPLM.

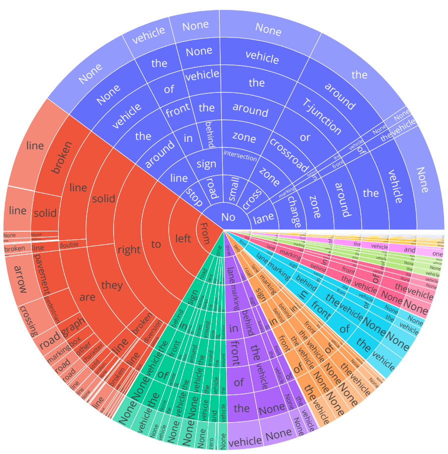


Figure 6. Word distribution in the HD map extracted captions of MAPLM. The figure is drawn based on 2,000 samples in the MAPLM dataset.

3.4. Evaluation Metrics

To test different VQA baselines for the MAPLM-QA task, we split the question-answer pairs into two types: Open QA and Fine-grained QA. Since the answer in Open QA is unstructured during annotation, we use rule-based metrics to evaluate the generated contents. To evaluate **LAN**, we extract the lane counting number from the output context and then calculate the correct ratio. The **DES** is defined based on the rule: if the **LAN** is predicted wrong, the **DES** will be 0; if **LAN** is predicted correct, the **DES** will be the correct ratio of each lane. Fine-grained QA can be considered as a multi-class classification problem with multiple options, thus they can be evaluated with the correct ratio as the accuracy metric. In addition to the evaluation of each item, we also propose to use two overall metrics: Frame-overall-accuracy (**FRM**) and Question-overall-accuracy (**QNS**). **FRM** is 1 if all Fine-grained QA and **LAN** are answered correctly for one frame, otherwise, it will be 0. **QNS** is the correct ratio of all questions.

$$\mathrm{DES} = \frac{1}{N}\sum_{k=1}^{N}(\mathrm{LAN}_k \cdot \frac{1}{M}\sum_{j=1}^{M}\mathrm{DES}_{k,j}) \tag{1}$$

$$\mathrm{FRM} = \frac{1}{N}\sum_{k=1}^{N}(\mathrm{SCN}_k \cdot \mathrm{QLT}_k \cdot \mathrm{LAN}_k \cdot \mathrm{INT}_k) \tag{2}$$

$$\mathrm{QNS} = \frac{1}{N}\frac{\sum_{k=1}^{N}\mathrm{SCN}_k + \mathrm{QLT}_k + \mathrm{LAN}_k + \mathrm{INT}_k}{4} \tag{3}$$

where $\mathrm{SCN}_k \in \{0,1\}$, $\mathrm{QLT}_k \in \{0,1\}$, $\mathrm{LAN}_k \in \{0,1\}$, $\mathrm{INT}_k \in \{0,1\}$, $\mathrm{DES}_{k,j} \in \{0,1\}$ are the binary result for related questions for one frame or one lane. N is the number of frames in the test set. M is the number of lanes for each frame.

4. Methodology - Baseline for MAPLM-QA

In this section, we present the baseline model, which serves as a multi-modal VLM developed for map and traffic scene comprehension in the domain of autonomous vehicles. The primary aim of this baseline model is to establish a standard for future research, enabling performance comparison for subsequent methods. It is important to note that our intent is not to surpass the performance of existing state-of-the-art multi-modal LLM approaches, but rather to facilitate consistent benchmarking. We first introduce the background and task definition of multi-modal traffic scene understanding in the context of autonomous driving and HD map analysis (Section 4.1). Then, we show the proposed multi-modal baseline model (Section 4.2). Finally, we introduce the two-stage pretraining and finetuning strategy for the MAPLM baseline (Section 4.3).

4.1. Task

The goal of multi-modal traffic scene understanding in the context of autonomous driving and HD map analysis is to align traffic and map context and driving perception such as panoramic 2D images and 3D LiDAR point cloud, and enhance downstream driving decision-making and explainable motion planning. The input of the task is multi-modal

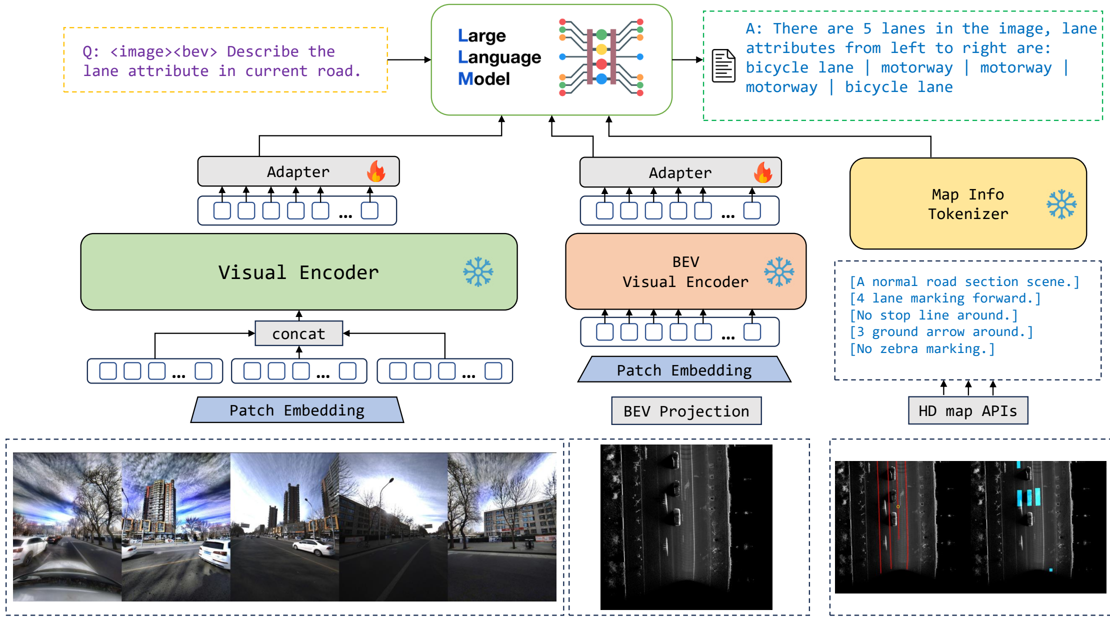


Figure 7. Schematic of the Baseline Instruction Tuning Model. The system ingests multi-modal inputs including panoramic 2D images, 3D LiDAR point clouds, and HD map contexts. Dual CLIP-based visual encoders are utilized to distill features from the images and point clouds respectively. These extracted features with the HD map info are integrated and processed by LLMs to synthesize coherent responses.

observations $O = \{X_v, X_{pc}, X_{hd}\}$ and the question X_q from question answer pair (X_q, Y). X_v, X_{pc}, X_{hd} is the panoramic image input, point cloud input, and HD map context extracted by other predefined segmentation or object detection models. X_q is the question input, while $\hat{Y}$ is the answer prediction. The multi-modal traffic scene understanding function F_θ can be formulated as:

$$\hat{Y} = F_\theta(O, X_q) \tag{4}$$

4.2. Baseline Framework Overview

As shown in Figure 7, MAPLM designed a simple baseline architecture using a multi-modal encoder and shared LLM decoder framework. The baseline model will be used for comparison with other state-of-the-art models.

Baseline Architecture. Following the idea from LLaVA [33], the MAPLM baseline model used patch embedding from CLIP to tokenize each panoramic 2D image into visual tokens. After concatenating tokens from different views into the input feature map, a pretrained CLIP visual encoder is used to extract joined features. We also generate a BEV representation from the 3D LiDAR point cloud. Each BEV representation is rotated in the direction of vehicle moving trajectories, and each pixel gray-scale value represents the reflection intensity of the local point cloud. The semantic information such as lane markings, ground signs, and zebra crossing in traffic scenes can be distinguished according to the light and dark changes of the reflection intensity. Then, several trainable projection matrix is used to align panoramic imaging tokens, and LiDAR point cloud BEV tokens into the text embedding space of the LLM.

Panoramic 2D image. For m input panoramic 2D images X_v^1, X_v^2, ..., X_v^m we use the same tokenizer ϕ_v from CLIP visual encoder (ViT-L/14-336) to embed them into tokens and then concatenate all tokens. The visual feature is extracted from CLIP's vision encoder and then the adaptor layer to map image features into the LLM's word embedding space:

$$Z_v = W_v \cdot f_v(\phi_v(X_v^1) \oplus \cdots \oplus \phi_v(X_v^m)), \quad Z_v \in R^{d \times k_v} \tag{5}$$

where $\oplus$ is the concatenation operation. f_v is the CLIP encoder. W_v is the weight of the adaptor layer. d is the dimension of the LLM embedding. k_v is the number of visual tokens.

LiDAR Point Cloud & BEV. The point cloud BEV image is tokenized by another CLIP's patch embedding ϕ_{bev} and then extracts point cloud features by CLIP visual encoder (ViT-L/14-336). A projection layer is used to map point cloud tokens into language tokens.

$$Z_{bev} = W_{bev} \cdot f_{bev}(\phi_{bev}(\mathrm{BEV}(X_{pc}))), \quad Z_{bev} \in R^{d \times k_{bev}} \tag{6}$$

where f_{bev} is the pretrained BEV visual encoder. W_{bev} is the projection matrix of the adaptor layer. d is the dimension of the LLM embedding. k_{bev} is the number of BEV visual tokens.

Question-Answer with Visual Features and HD Map Captions. For a sequence of length L, the autoregressive encoder in the LLM for generation answer is as follows:

$$P(\hat{Y}|Z_v, Z_{bev}, X_{hd}, X_q) = \prod_{i=1}^{L} P(y_i|Z_v, Z_{bev}, X_{hd}, X_{q,<i}, \hat{Y}_{<i}; \theta) \tag{7}$$

where $X_{q,<i}$ is all of the question tokens (the whole question) before y_i. $\hat{Y}_{<i}$ is all answer tokens before y_i. P is the conditional probability and θ is the trainable parameter in LLMs. In our experiment, we adopt Low-Rank Adaptation (LoRA) [25] to finetune the LLM models.

4.3. Training

Inspired by LLaVA [33] and InstructBLIP [12], MAPLM multi-modal baseline proposed a two-stage training strategy. The first stage is the pretraining of the CLIP visual encoder for BEV images. To balance modality coverage and pretraining efficiency, we merge and filter the 2M image-HD map information pairs to remove duplicated and similar road trajectories and finalize them to 510K image-text pairs. Then we used cleaned data to train the CLIP's visual encoder for BEV images. In the following experiment, we freeze the weights of both panoramic 2D images and BEV images' CLIP visual encoder.

In the second stage, we keep the CLIP weights frozen and focus on training both the panoramic 2D image and BEV image adaptor layers (projection layers) between the CLIP visual encoder and LLM. The adaptor layers for panoramic 2D images use the same initial weight from LLaVA [33]. The trainable parameters in the second stage are W_{bev}, W_v, and LoRA weight in the LLM.

5. Experiments and Results

Our experiment is designed to set up and test visual-language baselines and state-of-the-art methods on the proposed MAPLM-QA benchmark for all metrics.

5.1. Experimental Setting

We used the 510K image-text pairs data from MAPLM to pretrain the CLIP visual encoder. The MAPLM-QA dataset for instruction tuning is split into the train/validation/test set with 10775/1500/1500 frames, respectively, for all three tasks. For GPT-4V models, we used the official model API **gpt-4-vision-preview** (Access Date: Nov. 2023). All frames sent to GPT-4V include panoramic 2D images and one LiDAR BEV projection image. **0-shot** in Table 4 means no additional data from the training set are provided to GPT-4 in the input prompt. **5-shot** means 5 frames and QA annotations from the training set are provided to the input prompt as reference. For instruction tuning models, all models use LLaMA-2-7B [59] or Vicuna-7B [7] as the LLM. We pretrain and finetune them following the setups in Table 3 with 8 NVIDIA V100 GPUs in CLIP pretraining and 2 NVIDIA A100 GPUs for finetuning. Besides, to solve the class imbalance problem during baseline model finetuning, we randomly remove some questions based on their frequency of occurrence for each training epoch in the MAPLM-QA dataset.

Hyperparameters	Pretraining	Finetuning
batch size	16	2
learning rate	1e-4	1e-5
lr scheduler	cosine decay [35]	cosine decay [35]
lr warmup ratio	0.05	0.05
epoch	2	10
optimizer	AdamW [29]	AdamW [29]

Table 3. Hyperparameters setup. The rank in LoRA in the experiment is 128.

5.2. Results

Table 4 shows the quantitative comparison between zero-shot / few-shot GPT-4V [44] and instruction tuning-based VLMs. Table 5 is the ablation study to compare the GPT-4V (zero-shot) and baseline model's performance when using different modalities as input. GPT-4V [44] is the recently released cutting-edge VLM, which opens up new vistas for research and development. LLaVA [33] is an open-source VLM that showed strong multimodal chat abilities in various QA benchmarks [36]. After comparing these methods, we can observe that:

- Though it performed well in previous open-source datasets [65], GPT-4V demonstrated challenges in distinguishing the number of lanes in the MAPLM-QA test set. Sometimes, it generates incorrect responses to count lanes. These lane hallucination problems are likely due to the lack of relevant traffic scene reasoning information during model training [83].
- Initial weights of LLMs for visual instruction tuning will influence the multi-modal model's capability to learn traffic and map-related features.
- Using LoRA can improve the performance of visual instruction tuning for the MAPLM-QA benchmark.
- Both GPT-4V and baseline method's FRM and QNS increase when adding more modalities.

Method	Learning	Backbone	Open QA		Fine-grained QA			FRM(↑)	QNS(↑)
			LAN(↑)	DES((↑))	INT(↑)	QLT(↑)	SCN(↑)		
Random Select	-	-	21.00	-	16.73	25.20	15.27	0	19.55
GPT-4V [44]	0-shot	-	56.25	-	62.53	43.75	68.73	18.75	57.81
GPT-4V [44]	5-shot	-	58.32	-	74.33	53.18	69.57	20.18	60.94
LLaVA [33]	IT+LoRA	LLaMA-2-7B [59]	64.33	47.13	65.27	81.60	90.94	38.13	76.08
LLaVA [33]	IT+LoRA	Vicuna-7B [7]	75.40	64.89	77.53	82.40	95.53	52.27	82.72
Baseline	IT	LLaMA-2-7B [59]	59.67	47.03	75.87	77.47	92.53	36.27	76.38
Baseline	IT	Vicuna-7B [7]	72.93	62.75	78.40	82.27	94.93	50.53	82.13
Baseline	IT+LoRA	LLaMA-2-7B [59]	72.33	56.40	78.67	82.07	93.53	49.07	81.65
Baseline	IT+LoRA	Vicuna-7B [7]	**78.53**	**70.60**	**83.20**	**84.33**	**96.00**	**57.99**	**85.52**

Table 4. MAPLM QA Benchmark: Compare both GPT-4V [44] (Accessed Date: Nov, 2023) and state-of-the-art instruction tuning-based VLMs under MAPLM-QA benchmark. IT: Visual Instruction Tuning, LoRA: Low-Rank Adaptation [25]

Method	Modality		FRM	QNS
	image.	point cloud.		
GPT-4V [44]	✗	✓	12.57	53.28
	✓	✓	18.75	57.81
Baseline	✗	✓	45.47	80.25
	✓	✓	57.99	85.52

Table 5. Ablation study to evaluate the modalities as input. GPT-4V is under 0-shot setting

The result also proves that currently released LLMs can work on traffic and HD map data, however, it is still difficult to answer all questions for one frame correctly. Both GPT-4V and instruction tuning-based baseline can not achieve over 60% in FRM. Compared with GPT-4V, the instruction tuning-based baseline can answer well in lane counting. Furthermore, it is worth noting that the baseline model achieves 85.52% overall accuracy in answering all questions from MAPLM-QA (QNS) and 57.99% frame-level accuracy (FRM).

6. Discussion and Outlook

Map systems play a crucial role in autonomous driving navigation, with HD maps providing more refined information about the vehicles' operating environments [32, 37, 56, 70]. The integration of LLMs can significantly improve how HD maps are interpreted, leading to enhanced navigation precision and a deeper understanding of traffic scenarios. Our research introduces a new benchmark aimed at advancing this emerging field, advocating for the application of VLMs in aligning visual scenes and textual information within HD maps.

During our experimentation with MAPLM-QA, we identified a notable challenge: multi-modal LLMs trained with general domain knowledge often exhibit inaccuracies, such as lane misperceptions, particularly in scenarios not covered by existing open-source datasets. Although leveraging Reinforcement Learning from Human Feedback (RLHF) can mitigate these issues in the future, the time cost and safety concerns are still key limitations. In our paper, we explored visual instruction tuning as a potential solution. By integrating multi-modal inputs, the baseline model can significantly enhance performance in comprehending HD map scenes. Beyond understanding basic traffic elements, multi-modal LLMs hold the potential for higher-level reasoning about HD maps. In the future, traffic scene understanding in HD maps can be embedded with Mixture of Experts (MoE) [3, 21, 73] LLMs as additional API tools for current autonomous driving systems.

7. Conclusion

In this paper, we introduced MAPLM, a large-scale real-world vision-language dataset specifically designed for map and traffic scene understanding. In contrast to the existing dataset, MAPLM contains more data ensuring broad coverage of real-world scenarios, and can be used to solve the multi-modal data alignment among panoramic 2D images, 3D LiDAR point cloud, and text data extracted from HD maps. Our baseline model focused on using projected BEV images of 3D LiDAR point clouds and panoramic 2D images together with HD map descriptions to answer questions from MAPLM-QA. Our experimental results illuminate the need for further advancements in designing new multi-modal LLMs to fully leverage the dataset's potential. The dataset will be made fully available to the public to accelerate the progress of applying LLMs into this new field.

Acknowledgment

This work was supported by Tencent T lab. The contents of this paper only reflect the views of the authors, who are responsible for the facts and the accuracy of the data presented herein. The contents do not necessarily reflect the official views of Tencent.

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736, 2022. 2

[2] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015. 2

[3] Mikel Artetxe, Shruti Bhosale, Naman Goyal, Todor Mihaylov, Myle Ott, Sam Shleifer, Xi Victoria Lin, Jingfei Du, Srinivasan Iyer, Ramakanth Pasunuru, et al. Efficient large scale language modeling with mixtures of experts. *arXiv preprint arXiv:2112.10684*, 2021. 8

[4] Mehdi Azarafza, Mojtaba Nayyeri, Charles Steinmetz, Steffen Staab, and Achim Rettberg. Hybrid reasoning based on large language models for autonomous car driving. *arXiv preprint arXiv:2402.13602*, 2024. 1

[5] Daichi Azuma, Taiki Miyanishi, Shuhei Kurita, and Motoaki Kawanabe. Scanqa: 3d question answering for spatial scene understanding. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19129–19139, 2022. 2

[6] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11621–11631, 2020. 2, 3

[7] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*, 2(3):6, 2023. 7, 8

[8] DriveLM Contributors. Drivelm: Drive on language. `https://github.com/OpenDriveLab/DriveLM`, 2023. 2

[9] 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, 3

[10] Can Cui, Yunsheng Ma, Xu Cao, Wenqian Ye, Yang Zhou, Kaizhao Liang, Jintai Chen, Juanwu Lu, Zichong Yang, Kuei-Da Liao, et al. A survey on multimodal large language models for autonomous driving. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 958–979, 2024. 1, 2, 3

[11] Yaodong Cui, Shucheng Huang, Jiaming Zhong, Zhenan Liu, Yutong Wang, Chen Sun, Bai Li, Xiao Wang, and Amir Khajepour. Drivellm: Charting the path toward full autonomous driving with large language models. *IEEE Transactions on Intelligent Vehicles*, 2023. 1

[12] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning, 2023. 2, 7

[13] Thierry Deruyttere, Simon Vandenhende, Dusan Grujicic, Luc Van Gool, and Marie Francine Moens. Talk2car: Taking control of your self-driving car. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2088–2098, 2019. 2

[14] Vikrant Dewangan, Tushar Choudhary, Shivam Chandhok, Shubham Priyadarshan, Anushka Jain, Arun K Singh, Siddharth Srivastava, Krishna Murthy Jatavallabhula, and K Madhava Krishna. Talk2bev: Language-enhanced bird's-eye view maps for autonomous driving. *arXiv preprint arXiv:2310.02251*, 2023. 3

[15] Xinpeng Ding, Jianhua Han, Hang Xu, Wei Zhang, and Xiaomeng Li. Hilm-d: Towards high-resolution understanding in multimodal large language models for autonomous driving. *arXiv preprint arXiv:2309.05186*, 2023. 3

[16] Xinpeng Ding, Jinahua Han, Hang Xu, Xiaodan Liang, Wei Zhang, and Xiaomeng Li. Holistic autonomous driving understanding by bird's-eye-view injected multi-modal large models. *arXiv preprint arXiv:2401.00988*, 2024. 2

[17] Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. Carla: An open urban driving simulator. In *Conference on robot learning*, pages 1–16. PMLR, 2017. 3

[18] Mahdi Elhousni, Yecheng Lyu, Ziming Zhang, and Xinming Huang. Automatic building and labeling of hd maps with deep learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 13255–13260, 2020. 3

[19] Daocheng Fu, Wenjie Lei, Licheng Wen, Pinlong Cai, Song Mao, Min Dou, Botian Shi, and Yu Qiao. Limsim++: A closed-loop platform for deploying multimodal llms in autonomous driving. *arXiv preprint arXiv:2402.01246*, 2024. 3

[20] Daocheng Fu, Xin Li, Licheng Wen, Min Dou, Pinlong Cai, Botian Shi, and Yu Qiao. Drive like a human: Rethinking autonomous driving with large language models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 910–919, 2024. 3

[21] Trevor Gale, Deepak Narayanan, Cliff Young, and Matei Zaharia. MegaBlocks: Efficient Sparse Training with Mixture-of-Experts. *Proceedings of Machine Learning and Systems*, 5, 2023. 8

[22] Haoxiang Gao, Yaqian Li, Kaiwen Long, Ming Yang, and Yiqing Shen. A survey for foundation models in autonomous driving. *arXiv preprint arXiv:2402.01105*, 2024. 1

[23] Tao Gong, Chengqi Lyu, Shilong Zhang, Yudong Wang, Miao Zheng, Qian Zhao, Kuikun Liu, Wenwei Zhang, Ping Luo, and Kai Chen. Multimodal-gpt: A vision and language model for dialogue with humans. *arXiv preprint arXiv:2305.04790*, 2023. 2

[24] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating

the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913, 2017. 2

[25] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. 7, 8

[26] Yu Huang, Yue Chen, and Zhu Li. Applications of large scale foundation models for autonomous driving. *arXiv preprint arXiv:2311.12144*, 2023. 1, 3

[27] Yuichi Inoue, Yuki Yada, Kotaro Tanahashi, and Yu Yamaguchi. Nuscenes-mqa: Integrated evaluation of captions and qa for autonomous driving datasets using markup annotations. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 930–938, 2024. 2

[28] Jinkyu Kim, Anna Rohrbach, Trevor Darrell, John Canny, and Zeynep Akata. Textual explanations for self-driving vehicles. In *Proceedings of the European conference on computer vision (ECCV)*, pages 563–578, 2018. 2, 3

[29] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 7

[30] Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Jingkang Yang, and Ziwei Liu. Otter: A multi-modal model with in-context instruction tuning. *arXiv preprint arXiv:2305.03726*, 2023. 2

[31] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023. 2

[32] Qi Li, Yue Wang, Yilun Wang, and Hang Zhao. Hdmapnet: An online hd map construction and evaluation framework. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 4628–4634. IEEE, 2022. 8

[33] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*, 2023. 1, 2, 6, 7, 8

[34] Mingyu Liu, Ekim Yurtsever, Xingcheng Zhou, Jonathan Fossaert, Yuning Cui, Bare Luka Zagar, and Alois C Knoll. A survey on autonomous driving datasets: Data statistic, annotation, and outlook. *arXiv preprint arXiv:2401.01454*, 2024. 2

[35] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 7

[36] Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, 35:2507–2521, 2022. 2, 7

[37] Zhipeng Luo, Lipeng Gao, Haodong Xiang, and Jonathan Li. Road object detection for hd map: Full-element survey, analysis and perspectives. *ISPRS Journal of Photogrammetry and Remote Sensing*, 197:122–144, 2023. 8

[38] Yunsheng Ma, Can Cui, Xu Cao, Wenqian Ye, Peiran Liu, Juanwu Lu, Amr Abdelraouf, Rohit Gupta, Kyungtae Han, Aniket Bera, et al. Lampilot: An open benchmark dataset for autonomous driving with language model programs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 3

[39] Srikanth Malla, Chiho Choi, Isht 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, 3

[40] Jiageng Mao, Yuxi Qian, Hang Zhao, and Yue Wang. Gpt-driver: Learning to drive with gpt. *arXiv preprint arXiv:2310.01415*, 2023. 3

[41] Jiageng Mao, Junjie Ye, Yuxi Qian, Marco Pavone, and Yue Wang. A language agent for autonomous driving. *arXiv preprint arXiv:2311.10813*, 2023. 3

[42] Aboli Marathe, Deva Ramanan, Rahee Walambe, and Ketan Kotecha. Wedge: A multi-weather autonomous driving dataset built from generative vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3317–3326, 2023. 2

[43] 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. *arXiv preprint arXiv:2312.14115*, 2023. 2, 3

[44] OpenAI. Gpt-4v(ision) system card. `https://openai.com/research/gpt-4v-system-card`, 2023. 2, 3, 7, 8

[45] David Pannen, Martin Liebner, Wolfgang Hempel, and Wolfram Burgard. How to keep hd maps for automated driving up to date. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 2288–2294. IEEE, 2020. 3

[46] Baolin Peng, Chunyuan Li, Pengcheng He, Michel Galley, and Jianfeng Gao. Instruction tuning with gpt-4. *arXiv preprint arXiv:2304.03277*, 2023. 2

[47] 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. *arXiv preprint arXiv:2305.14836*, 2023. 2

[48] Yujia Qin, Shengding Hu, Yankai Lin, Weize Chen, Ning Ding, Ganqu Cui, Zheni Zeng, Yufei Huang, Chaojun Xiao, Chi Han, et al. Tool learning with foundation models. *arXiv preprint arXiv:2304.08354*, 2023. 1, 2

[49] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 2

[50] Enna Sachdeva, Nakul Agarwal, Suhas Chundi, Sean Roelofs, Jiachen Li, Mykel Kochenderfer, Chiho Choi, and Behzad Dariush. Rank2tell: A multimodal driving dataset for joint importance ranking and reasoning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 7513–7522, 2024. 2

[51] Heiko G Seif and Xiaolong Hu. Autonomous driving in the icity—hd maps as a key challenge of the automotive industry. *Engineering*, 2(2):159–162, 2016. 3

[52] Hao Shao, Yuxuan Hu, Letian Wang, Steven L Waslander, Yu Liu, and Hongsheng Li. Lmdrive: Closed-loop end-to-end driving with large language models. *arXiv preprint arXiv:2312.07488*, 2023. 3

[53] Yongliang Shen, Kaitao Song, Xu Tan, Dongsheng Li, Weiming Lu, and Yueting Zhuang. Hugginggpt: Solving ai tasks with chatgpt and its friends in huggingface. *arXiv preprint arXiv:2303.17580*, 2023. 2

[54] Chonghao Sima, Katrin Renz, Kashyap Chitta, Li Chen, Hanxue Zhang, Chengen Xie, Ping Luo, Andreas Geiger, and Hongyang Li. Drivelm: Driving with graph visual question answering. *arXiv preprint arXiv:2312.14150*, 2023. 2

[55] Ishika Singh, Valts Blukis, Arsalan Mousavian, Ankit Goyal, Danfei Xu, Jonathan Tremblay, Dieter Fox, Jesse Thomason, and Animesh Garg. Progprompt: Generating situated robot task plans using large language models. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 11523–11530. IEEE, 2023. 1

[56] Kun Tang, Xu Cao, Zhipeng Cao, Tong Zhou, Erlong Li, Ao Liu, Shengtao Zou, Chang Liu, Shuqi Mei, Elena Sizikova, et al. Thma: Tencent hd map ai system for creating hd map annotations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 15585–15593, 2023. 3, 8

[57] Xiaoyu Tian, Junru Gu, Bailin Li, Yicheng Liu, Chenxu Hu, Yang Wang, 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. 3

[58] Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal llms. *arXiv preprint arXiv:2401.06209*, 2024. 1, 2

[59] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023. 7, 8

[60] Wenhai Wang, Jiangwei Xie, ChuanYang Hu, Haoming Zou, Jianan Fan, Wenwen Tong, Yang Wen, Silei Wu, Hanming Deng, Zhiqi Li, et al. Drivemlm: Aligning multi-modal large language models with behavioral planning states for autonomous driving. *arXiv preprint arXiv:2312.09245*, 2023. 3

[61] Yixuan Wang, Ruochen Jiao, Chengtian Lang, Sinong Simon Zhan, Chao Huang, Zhaoran Wang, Zhuoran Yang, and Qi Zhu. Empowering autonomous driving with large language models: A safety perspective. *arXiv preprint arXiv:2312.00812*, 2023. 1

[62] Wayve. Lingo-1: Exploring natural language for autonomous driving. https://wayve.ai/thinking/lingo-natural-language-autonomous-driving/, 2023. 3

[63] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*, 2022. 1

[64] 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. 3

[65] Licheng Wen, Xuemeng Yang, Daocheng Fu, Xiaofeng Wang, Pinlong Cai, Xin Li, Tao Ma, Yingxuan Li, Linran Xu, Dengke Shang, et al. On the road with gpt-4v (ision): Early explorations of visual-language model on autonomous driving. *arXiv preprint arXiv:2311.05332*, 2023. 1, 2, 4, 7

[66] Kelvin Wong, Yanlei Gu, and Shunsuke Kamijo. Mapping for autonomous driving: Opportunities and challenges. *IEEE Intelligent Transportation Systems Magazine*, 13(1):91–106, 2020. 3

[67] Chenfei Wu, Shengming Yin, Weizhen Qi, Xiaodong Wang, Zecheng Tang, and Nan Duan. Visual chatgpt: Talking, drawing and editing with visual foundation models. *arXiv preprint arXiv:2303.04671*, 2023. 2

[68] Dongming Wu, Wencheng Han, Tiancai Wang, Yingfei Liu, Xiangyu Zhang, and Jianbing Shen. Language prompt for autonomous driving. *arXiv preprint*, 2023. 2

[69] Penghao Wu and Saining Xie. V*: Guided visual search as a core mechanism in multimodal llms. *arXiv preprint arXiv:2312.14135*, 2023. 2

[70] Ziyang Xie, Ziqi Pang, and Yu-Xiong Wang. Mv-map: Offboard hd-map generation with multi-view consistency. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8658–8668, 2023. 8

[71] Li Xu, He Huang, and Jun Liu. Sutd-trafficqa: A question answering benchmark and an efficient network for video reasoning over traffic events. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9878–9888, 2021. 2, 3

[72] Zhenhua Xu, Yujia Zhang, Enze Xie, Zhen Zhao, Yong Guo, Kenneth KY Wong, Zhenguo Li, and Hengshuang Zhao. Drivegpt4: Interpretable end-to-end autonomous driving via large language model. *arXiv preprint arXiv:2310.01412*, 2023. 3

[73] Fuzhao Xue, Zian Zheng, Yao Fu, Jinjie Ni, Zangwei Zheng, Wangchunshu Zhou, and Yang You. Openmoe: An early effort on open mixture-of-experts language models. *arXiv preprint arXiv:2402.01739*, 2024. 8

[74] Ke Yang, Jiateng Liu, John Wu, Chaoqi Yang, Yi R Fung, Sha Li, Zixuan Huang, Xu Cao, Xingyao Wang, Yiquan Wang, et al. If llm is the wizard, then code is the wand: A survey on how code empowers large language models to serve as intelligent agents. *arXiv preprint arXiv:2401.00812*, 2024. 1

[75] Senqiao Yang, Jiaming Liu, Ray Zhang, Mingjie Pan, Zoey Guo, Xiaoqi Li, Zehui Chen, Peng Gao, Yandong Guo, and Shanghang Zhang. Lidar-llm: Exploring the potential of large language models for 3d lidar understanding. *arXiv preprint arXiv:2312.14074*, 2023. 3

[76] 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, 2

[77] Zhengyuan Yang, Linjie Li, Kevin Lin, Jianfeng Wang, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. The dawn of lmms: Preliminary explorations with gpt-4v (ision). *arXiv preprint arXiv:2309.17421*, 2023. 2

[78] Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. A survey on multimodal large language models. *arXiv preprint arXiv:2306.13549*, 2023. 1

[79] Zhenfei Yin, Jiong Wang, Jianjian Cao, Zhelun Shi, Dingning Liu, Mukai Li, Lu Sheng, Lei Bai, Xiaoshui Huang, Zhiyong Wang, et al. Lamm: Language-assisted multi-modal instruction-tuning dataset, framework, and benchmark. *arXiv preprint arXiv:2306.06687*, 2023. 2

[80] Zhou Yu, Dejing Xu, Jun Yu, Ting Yu, Zhou Zhao, Yueting Zhuang, and Dacheng Tao. Activitynet-qa: A dataset for understanding complex web videos via question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 9127–9134, 2019. 2

[81] Jianhao Yuan, Shuyang Sun, Daniel Omeiza, Bo Zhao, Paul Newman, Lars Kunze, and Matthew Gadd. Rag-driver: Generalisable driving explanations with retrieval-augmented incontext learning in multi-modal large language model. *arXiv preprint arXiv:2402.10828*, 2024. 3

[82] Yihan Zeng, Chenhan Jiang, Jiageng Mao, Jianhua Han, Chaoqiang Ye, Qingqiu Huang, Dit-Yan Yeung, Zhen Yang, Xiaodan Liang, and Hang Xu. Clip2: Contrastive language-image-point pretraining from real-world point cloud data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15244–15253, 2023. 2

[83] Chenhui Zhang and Sherrie Wang. Good at captioning, bad at counting: Benchmarking gpt-4v on earth observation data. *arXiv preprint arXiv:2401.17600*, 2024. 2, 7

[84] Shengyu Zhang, Linfeng Dong, Xiaoya Li, Sen Zhang, Xiaofei Sun, Shuhe Wang, Jiwei Li, Runyi Hu, Tianwei Zhang, Fei Wu, et al. Instruction tuning for large language models: A survey. *arXiv preprint arXiv:2308.10792*, 2023. 2

[85] Chao Zheng, Xu Cao, Kun Tang, Zhipeng Cao, Elena Sizikova, Tong Zhou, Erlong Li, Ao Liu, Shengtao Zou, Xinrui Yan, et al. High-definition map automatic annotation system based on active learning. *AI Magazine*, 44(4):418–430, 2023. 3

[86] Xingcheng Zhou, Mingyu Liu, Bare Luka Zagar, Ekim Yurtsever, and Alois C Knoll. Vision language models in autonomous driving and intelligent transportation systems. *arXiv preprint arXiv:2310.14414*, 2023. 1, 2

[87] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023. 2