

TB-Bench: Training and Testing Multi-Modal AI for Understanding Spatio-Temporal Traffic Behaviors from Dashcam Images/Videos

Korawat Charoenpitaks^{1,*}, Van-Quang Nguyen^{2,*}, Masanori Suganuma¹, Kentaro Arai³,
Seiji Totsuka³, Hiroshi Ino³, Takayuki Okatani^{1,2,*}
¹Tohoku University, ²RIKEN AIP, ³DENSO CORPORATION

Abstract

The application of Multi-modal Large Language Models (MLLMs) in Autonomous Driving (AD) faces significant challenges due to their limited training on traffic-specific data and the absence of dedicated benchmarks for spatiotemporal understanding. This study addresses these issues by proposing TB-Bench, a comprehensive benchmark designed to evaluate MLLMs on understanding traffic behaviors across eight perception tasks from ego-centric views. We also introduce vision-language instruction tuning datasets, TB-100k and TB-250k, along with simple yet effective baselines for the tasks. Through extensive experiments, we show that existing MLLMs underperform in these tasks, with even a powerful model like GPT-4o achieving less than 35% accuracy on average. In contrast, when fine-tuned with TB-100k or TB-250k, our baseline models achieve average accuracy up to 85%, significantly enhancing performance on the tasks. Additionally, we demonstrate performance transfer to another driving benchmark by co-training a model on the other driving benchmark dataset with our proposed dataset. The benchmark, datasets, and code will be available at <https://github.com/TB-AD/TB-Bench>.

1. Introduction

Multi-modal Large Language Models (MLLMs) have been increasingly applied to Autonomous Driving (AD), particularly for predicting risks and planning actions based on images or videos from in-vehicle cameras. Notably, MLLMs have demonstrated their effectiveness in the international competitions like Autonomous Grand Challenge [37] and in specific tasks such as traffic sign detection [58]. However, two major challenges remain.

First, current MLLMs, ranging from proprietary models like GPT-4o [2] and Gemini [42] to open-source models

*Corresponding authors: {korawat, okatani}@vision.is.tohoku.ac.jp, quang.nguyen.jz@riken.jp

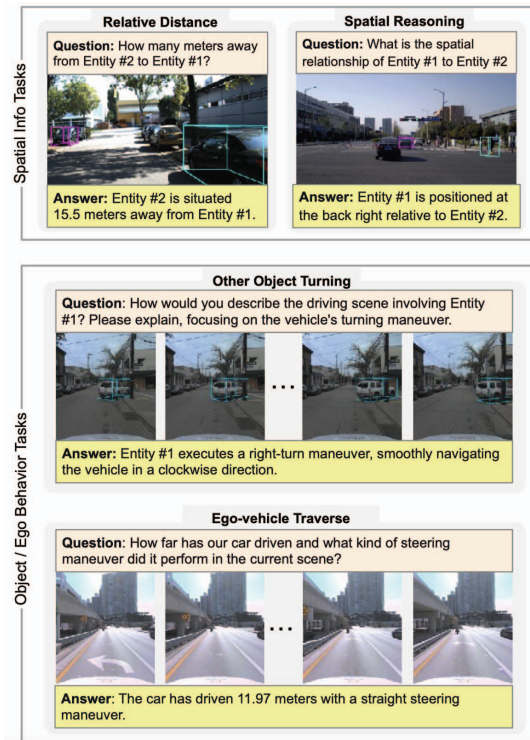


Figure 1. Examples of four tasks from TB-Bench; additional task examples are provided in the supplementary material.

like LLaVA [26], are not optimized for dashcam images or traffic scenes. These models are primarily trained on vast amounts of web-based text and image-text pairs, with minimal traffic-specific data, limiting their effectiveness in AD scenarios. To improve the generalizability of MLLMs, incorporating high-quality domain-specific datasets into the pre-training data is crucial, as shown in [21, 57].

Second, it lacks a dedicated benchmark for evaluating MLLMs' abilities in spatiotemporal understanding tasks, given their capabilities in vision-centric tasks are still developing. While these models are designed to handle diverse vision-language tasks, they struggle with complex visual understanding, such as spatial reasoning and object relation-

ships [43]. Even in common domains unrelated to traffic scenes, there are insufficient benchmarks (e.g., Cambrian-1 [44], V-star benchmark [50]). Given that AD requires sophisticated geometric and spatiotemporal understanding to capture dynamic vehicle-entity interactions, high-quality dedicated benchmarks are needed.

While MLLMs are increasingly applied in AD, the aforementioned challenges remain insufficiently addressed. Recent research has primarily focused on using pretrained MLLMs derived from web data for specific AD tasks, without thoroughly investigating these challenges. Another issue is to determine which AD tasks across different stages and levels should be addressed by MLLMs. Within the “perception, prediction, and planning” framework, the question becomes: which stages should MLLMs handle?

Focusing solely on the perception stage and its associated tasks, it may not necessarily appear optimal to use MLLMs. Established technologies like LIDAR and CV methods such as object detection and visual odometry can accurately capture the vehicle’s position and spatiotemporal relationships in Euclidean space.

However, when considering the use of MLLMs (or LLMs) in later stages, relying on such “Euclidean geometrically accurate” information in the earlier stage might not be optimal. Firstly, it is unclear how to input this information into LLMs. Moreover, achieving advanced understanding in later stages may require information representations specifically suited for MLLMs to perform higher-level tasks. This suggests the need for MLLM involvement from the perception stage.

This study adopts this perspective, aiming to involve MLLMs in the perception stage by expressing spatiotemporal traffic scene tasks in natural language text. The underlying conjecture, as mentioned above, is that this approach could be important for prediction and planning in later stages by MLLMs.

To address the challenge of lacking dedicated benchmarks, we introduce TB-Bench, one of the first comprehensive benchmarks specifically designed to evaluate MLLM’s understanding of traffic behaviors. This benchmark assesses MLLM’s capabilities to perform perception tasks based on dashcam images or videos from the ego-centric views of vehicles, including determining the spatial position or orientation of other vehicles and interpreting the behaviors of both ego-vehicles and surrounding traffic. Compared to existing benchmarks, TB-Bench encompasses a wider array of eight distinct perception tasks,¹ each corresponding to a typical driver maneuver. Figure 1 shows examples of several tasks. To ensure consistent evaluation across a diverse range of MLLMs, we employ a straightforward pro-

cedure. Specifically, we pair questions with images or video clips, requiring an MLLM to respond in plain text. Performance of the MLLM is then assessed by measuring response accuracy.

To address the challenge of insufficient training data for AD perception tasks, we introduce a high-quality dataset focused on traffic behavior understanding from ego-centric views. This dataset aligns with the task design of TB-Bench and is used for vision-language instruction tuning (VLIT) of MLLMs. We generate high-quality question-and-answer pairs using samples from established datasets such as KITTI, ONCE, and Argoverse 2. Specifically, we constructed **TB-Bench**, which consists of 2,000 manually created samples, along with two versions of training datasets: **TB-250k**, containing 250,000 samples, and **TB-100k**, a more balanced version with 100,000 samples.

In addition to evaluating existing MLLMs, we introduce a generic framework that serves as a strong baseline for our tasks, consisting of three standard components: a pretrained vision encoder, a multi-modal connector, and a pretrained LLM. The vision encoder extracts visual representations from inputs with varying number of frames, while the connector projects these embeddings into the LLM’s embedding space, finally the LLM generates task-specific responses on our benchmark. This lightweight model is designed for efficient fine-tuning on our proposed dataset(s).

Using TB-Bench to evaluate popular proprietary models (GPT-4o and Gemini) and various state-of-the-art open-source MLLMs (LLaVA, Bunny, and InternVL), we find that none of these models excels across all traffic behavior understanding tasks. On average, the open-source models underperform random guessing, while proprietary models achieve only slightly better results, with average accuracy below 35%. In contrast, when fine-tuned on TB-100k or TB-250k, our proposed baseline models demonstrate strong performance across all tasks, with average accuracy ranging from 77% to 85%. This highlights the effectiveness of our dataset in enhancing MLLM traffic behavior understanding.

Overall, our contributions are fourfold: 1) we introduce TB-Bench, a benchmark for assessing MLLMs on eight perception tasks of traffic behavior understanding; 2) we present the VLIT datasets (TB-100k and TB-250k) for the tasks, along with a generic baseline; 3) we conduct extensive experiments demonstrating the performance gap between existing MLLMs and the fine-tuned baselines; and 4) we show that our proposed dataset would be used as part of co-training datasets to generalize and improve the performance on other driving benchmarks, such as BDD-X [19].

2. Related Work

A summary of existing studies and benchmarks across various AD tasks is presented in Table 1.

¹Eight proposed tasks: Relative Distance, Spatial Reasoning, Orientation Reasoning, Other Lane to Ego, Other Lane Changing, Other Turning, Ego Turning, and Ego Traverse Distance.

Table 1. Summary of existing studies and benchmarks across AD tasks (brackets indicate tasks involving planning).

Benchmarking	Visual Data Modality	Perception (Planning) Tasks	Abbreviation	Meaning
Standalone Task in AD			OD	2D & 3D Object Detection
DRAMA [29]	single-image	PER, REA	OT	2D & 3D Object Tracking
Rank2Tell [38]	single-image	PER, REA, LANE, TLS	D	Depth Estimation
BDD-X [19]	multi-frame	PER, (AC)	OBJ	Object Existence, Class, etc.
BDD-OIA [52]	single-image	PER	KNOW	World Knowledge
TrafficQA [51]	multi-frame	PER, PRED, REA	LOC	Location or Coordinate
LingoQA [31]	multi-frame	PER, PRED, REA	LANE	Road, Lane, Intersection, etc
NuScenes-QA [35]	multi-view	OBJ, SP	PER	General Perception
NuScenes-MQA [17]	multi-view	OBJ, RD, OD	PRED	General Prediction
MAPLM-QA [5]	multi-view, BEV-image	LANE	PLAN	General Planning
DriveLM [40]	single-image	PER, PRED, (PLAN)	REA	General Reasoning
Benchmark			TLS	Traffic Light or Sign
SpatialRGPT [7]	single-image	RD, SR, OR	AC	Action Category
SEED [20]	multi-image, multi-frame	PER, PRED, REA, AR	AR	General Action Recognition
MVBench [23]	multi-frame	PER, PRED, REA, LOC, AR	RD	Relative Distance
MME [11] / MME-Realworld [59]	single-image	PER, PRED	SR	Spatial Reasoning
MMMU [55]	multi-image	PER, REA, KNOW	OR	Orientation Reasoning
ELM [60]	multi-frame	PER, PRED, TLS, OD, OT, AR, (PLAN)	EGO-LANE	Other Lane to Ego-vehicle
Cambrian-1 [44]	single-image	RD, SR, D	OBJ-LANE	Other Lane Changing
OpenEQA [28]	multi-frame	OBJ, SR , KNOW, LOC, REA	OBJ-TURN	Other Turning
TB-Bench (Ours)	single-image, multi-frame	RD, SR, OR, EGO-LANE, OBJ-LANE, OBJ-TURN, EGO-TURN, EGO-TRA	EGO-TURN	Ego Turning
			EGO-TRA	Ego Traverse Distance

2.1. Autonomous Driving Tasks

The majority of evaluations in the AD field are focused on either end-to-end driving systems, open-loop planning, or standalone task schemes, such as single-round visual question answering (VQA) or captioning. Traditionally, the AD framework consists of perception, prediction, and planning tasks [33], although slight variations exist, i.e., predicting intention-level outputs instead of trajectories [43].

Generally, perception tasks in end-to-end driving systems are mainly auxiliary tasks, consisting of all available supervision signals provided based on the data source. For example, NuScene [4] provides BEV information, segmentation labels, and more. Consequently, multi-task learning is applied to these tasks, such as object detection, tracking, and segmentation. This approach is consistent across recent similar AD planning datasets, whether in open-loop or simulation scenarios. Occasionally, pretrained VL models are utilized to enhance these modules.

Other popular traffic planning datasets are KITTI [14], ONCE [30], Waymo Open [41], and Argoverse2 [46], which are inherently similar to NuScene in characteristics.

Pretrained VL models are commonly known for their excellence in scene understanding, details, and visual cues. Still, it shows limitations in spatial grounding and reasoning [43]. In detail, most standalone task schemes focus on perception tasks, which include general event VQA [31, 51], environment and weather conditions, traffic signals, and lane information [5, 45]. These tasks also encompass critical object detection [29, 38] or tracking in various forms, such as bounding box coordinates [43], region proposals [9, 52], 2D [48], and 3D [49] language-guided object tracking, as well as scene analysis that includes attributes or motion of objects like size, position, direction, distance, spa-

tial position relationships [35], and orientation [7]. In particular, a comprehensive driving task integrates language with perception and prediction, features multi-choice VQA evaluation [59], and includes planning questions [40]. In the prediction tasks, all previous perception inputs are used to predict the object’s future trajectory, such as parking or moving, and interactions with the ego-vehicle. In the planning stage, it involves combining prior information to generate actions, decision descriptions [52], and trajectory waypoints [40, 43].

2.2. MLLMs and Benchmarks

VL pre-training and foundation models started with learning from a broader source of supervision, specifically raw text at an internet scale [36], enabling zero-shot transfer of the model to downstream tasks. Notably, approaches attempting to connect VL pre-training to existing LLMs, referred to as MLLMs [22], enable capabilities similar to those of LLMs, such as image-to-text generation, improved via instruction tuning and in-context learning capabilities. Current frontier families of MLLMs, such as LLaVA [26], VILA [24], and InternVL [6], utilize a similar architectural paradigm: vision encoder, multi-modal projector, and LLM connected in sequence. Despite some early work attempting resampler techniques like Q-Former [8], all state-of-the-art models use simpler linear layers with scaling to higher resolutions, focusing on higher quality VLIT instead. Another line of studies works on lightweight versions of MLLMs, optimizing for more informative, condensed training data and design choices [15, 39]. The latest MLLMs focus on simultaneously tackling multi-image, multi-frame (video), multi-view (3D), and multi-patch (single-image) scenarios, which show emergent capabilities and enhance overall per-

formance [21]. Nevertheless, it is a standard paradigm for MLLMs to evaluate on multiple general benchmarks, aiming to achieve overall performance.

The existing benchmarks, which refer to MLLM benchmarks, aim to comprehensively evaluate various dimensions, but there is no standardized taxonomy for benchmark design. General benchmarks in the VL space started with simple perception-oriented tasks [11], followed by multi-frame benchmarks [20, 23] with action recognition and VL knowledge-based reasoning [55]. Spatial or vision-centric benchmarks [7, 44] are becoming more relevant to address previously claimed weaknesses. Then, specialized benchmarks gained more attention, introducing tasks from different domains, such as robotics [28] and AD [40, 59]. In this case, there is still a lack of studies covering simple yet very important skills and behaviors in the AD context.

3. Benchmark Design

3.1. Task Design

We generate question-answer pairs in VQA format, where the model takes an image or video with a question as input and produces a corresponding answer. Both question and answer are expressed in a single sentence of free-form text.

To achieve the above goal, we consider multiple types of Q&A pairs, each linked to a specific driver’s maneuver behavior. We refer to the Pre-crash Scenarios typology from the National Automotive Sampling System (NASS) variables [32], which are also utilized in the CARLA simulator [10]. This typology includes a total of 65 pre-crash scenarios, categorized into nine accident types². Each scenario is described in the format of ‘an accident type: a detailed scenario.’ For example, the ‘lane change’ accident type includes scenarios like ‘one vehicle passing while another is turning.’ See the supp. material for the full list of scenarios.

Focusing on typical maneuver behaviors derived from NASS scenarios, we have identified eight distinct Q&A types, referred to as ‘tasks,’ as shown in Table 2. Some tasks require numerical outputs (e.g., ‘distance in meters’), while others require discrete classes (e.g., ‘back,’ ‘back left,’ etc.). It is important to note that the models are expected to provide these outputs in their natural language responses. Fig. 1 presents examples for four of the eight tasks, each of which consists of input image(s) accompanied by a question and a ground-truth answer. The visual input is either a single image or multiple images (up to eight), depending on the task, as will be explained later.

3.2. Referencing Entities

Some tasks require the model to determine the spatial position or orientation of other vehicles, as shown in Fig. 1.

²The accident types are Animal, Off-road, Pedalcyclist, Pedestrian, Backing, Lane Change, Opposite Direction, Rear-end, and Crossing-paths.

Table 2. **Tasks and Concepts Addressed in Each.** ‘Classes’ column indicates the types of outputs, i.e., the number of discrete classes or numerical outputs (indicated by $\mathcal{R}$); ‘Orientation Reasoning’ task contains both output types.

Task Type	Abstract Concepts	Classes
Spatial Information:		
Relative Distance	distance in meters	$\mathcal{R}$
Spatial Reasoning	back, back left, back right, front, front left, front right	6
Orientation Reasoning	opposite, perpendicular, similar, and degrees	3/ $\mathcal{R}$
Object Behavior:		
Other Lane to Ego-Vehicle	front lane, front left lane, front right lane, oncoming traffic lane	4
Other Lane Changing	left lane change, no change, right lane change	3
Other Turning	go straight, left turn, right turn	3
Ego Behavior:		
Ego Turning	go straight, left turn, right turn	3
Ego Traverse Distance	distance traveled in meters	$\mathcal{R}$

When multiple vehicles are present in a scene, it is essential to distinguish between them in both the questions and answers. One approach is to describe the vehicle by its attributes, such as “black compact sedan,” but this can pose challenges in ensuring the model accurately identifies and differentiates similar objects using such descriptions. To avoid these complications and focus on evaluating the model’s spatial understanding, we label each target traffic entity as ‘Entity # n ’ in the questions and answers, where n corresponds to its index in the input image(s); see examples in the upper part of Fig. 1. To identify these entities, we draw colored three-dimensional bounding boxes (BBs) directly in the input image(s), using a consistent color for each entity index n throughout the dataset. Specifically, we use cyan and magenta BBs for ‘Entity #1’ and ‘Entity #2,’ respectively. Our dataset includes up to two entities per scene, i.e., $n = 1$ or 2. An additional advantage of this method is that it requires minimal instruction tuning or even no extra learning for MLLMs to adapt. It is compatible with multi-view, multi-frame, and multi-scale modalities, as demonstrated in AnyRes [25], UniRes [57], and Interleave [21].

3.3. Evaluation

Our benchmark requires MLLMs to generate plain text outputs. Since the goal is to evaluate the spatiotemporal understanding capabilities of MLLMs, their output accuracy should be evaluated using methods suited to this goal.

The questions in the dataset are broadly classified into two categories based on the type of answers expected. One category includes questions about positional relationships or orientation, with typical answers like “positioned at the back right” or “a right-turn maneuver.” The other category involves questions requiring numerical answers, such as “is situated 15.53 meters away.”

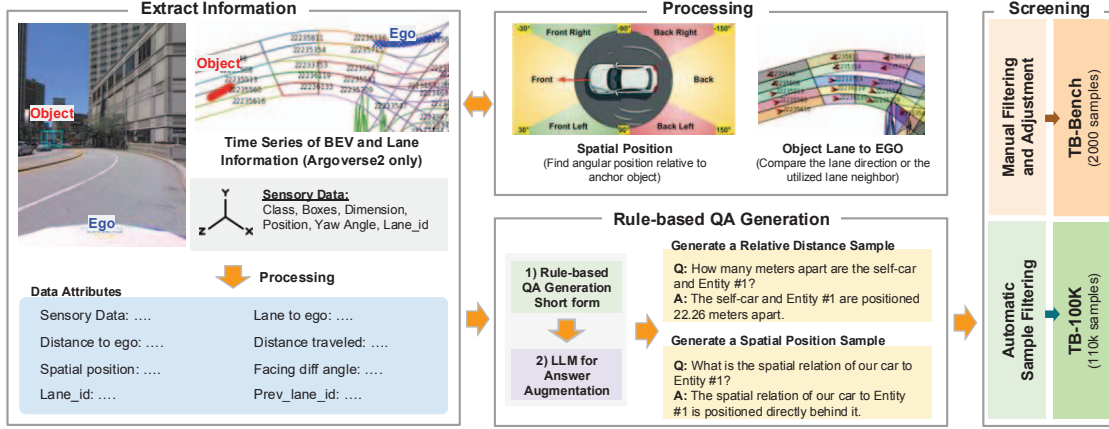


Figure 2. **Overview of Data Generation Pipeline.** Left: Sensory data is processed into higher-level attributes. Middle-Top: Spatial positioning and lane orientation relative to the ego-vehicle are determined. Middle-Bottom: Q&A samples are generated using rules and LLM augmentation. Right: Data is filtered and refined for the final dataset.

For the first category of Q&A, keywords are manually selected for each task or ground truth answer, and their presence in the output text is identified using rule-based methods (i.e., regular expressions). For the second category, the predicted value is compared to the correct answer, and if the difference falls within a specified range, the prediction is considered correct; otherwise, it is deemed incorrect. In the experiments, thresholds are set such that a difference within 25% of the correct value is considered acceptable for distance, and a difference within 15 degrees is acceptable for angle. Refer to the supp. material for more details.

4. Generation of VQA Data

4.1. Outline

To generate Q&A pairs for the eight tasks mentioned, we repurpose existing datasets, specifically KITTI [13], ONCE [30], and Argoverse2 [46]. These datasets are originally designed for studying object detection, localization, and tracking in three-dimensional space, providing detailed three-dimensional geometry of traffic entities. KITTI and ONCE, in particular, offer object class information and 3D bounding boxes for each traffic entity, including their position, dimensions, and yaw angle. Argoverse2 further enriches this with lane information relative to the ego vehicle.

To align with the task design mentioned (Table 2), the quantities provided by these datasets, mostly represented in the Euclidean space, are converted into abstract concepts, such as six discrete angles between two vehicles (e.g., front right, back left, etc.), lanes relative to the ego-car (i.e., front left lane, oncoming lane) and lane changing.

For the first three tasks—‘Relative Distance,’ ‘Spatial Reasoning,’ and ‘Orientation Reasoning’—we generate Q&A pairs using samples from KITTI and ONCE, as these tasks do not require lane information from the ego vehicle

Table 3. **Statistics of TB-Bench, TB-100k, and TB-250K.** Source datasets: K (KITTI), O (ONCE), Arv2 (Argoverse2).

Task Type	Sources/ Frames	TB- Bench	TB- 250k	TB- 100k
Spatial Information:				
Relative Distance	[K, O]/1	250	35k	10k
Spatial Reasoning	[K, O]/1	250	70k	30k
Orientation Reasoning	[K, O]/1	250	70k	30k
Object Behavior:				
Other Lane to Ego	[Arv2]/8	250	50k	20k
Other Lane Changing	[Arv2]/8	250	1.5k	1.5k
Other Turning	[Arv2]/8	250	1.5k	1.5k
Ego Behavior:				
Ego Turning	[Arv2]/8	250	1.5k	1.5k
Ego Traverse Distance	[Arv2]/8	250	25k	15.5k
Total		2000	254k	110k

or others. Since these tasks can be performed using a single image, we utilize a static dashcam image as the visual input. For the remaining tasks—‘Other Lane to Ego,’ ‘Other Lane Changing,’ ‘Other Turning,’ ‘Ego Turning,’ and ‘Ego Traverse Distance’—which require lane information and a multi-frame source, we generate Q&A pairs using Argoverse2. Given that these tasks involve temporal changes, we extract eight image frames from the ‘long scenario’ sequences in the dataset for each Q&A pair³, using these sequences as the visual input for models.

After generating the data automatically, we conduct a manual screening process. Based on the extent of screening, the data is organized into three distinct datasets. One dataset, comprising 2,000 samples, is designated for evaluation purposes, which we will refer to as ‘benchmark’ in this paper. These samples undergone thorough manual inspection, removing low-quality samples and ensuring an equal

³Each ‘long scenario’ sequence in the dataset is 15 seconds long. From these, we extract 1.6-second clips, consisting of eight images captured at 0.2-second intervals.

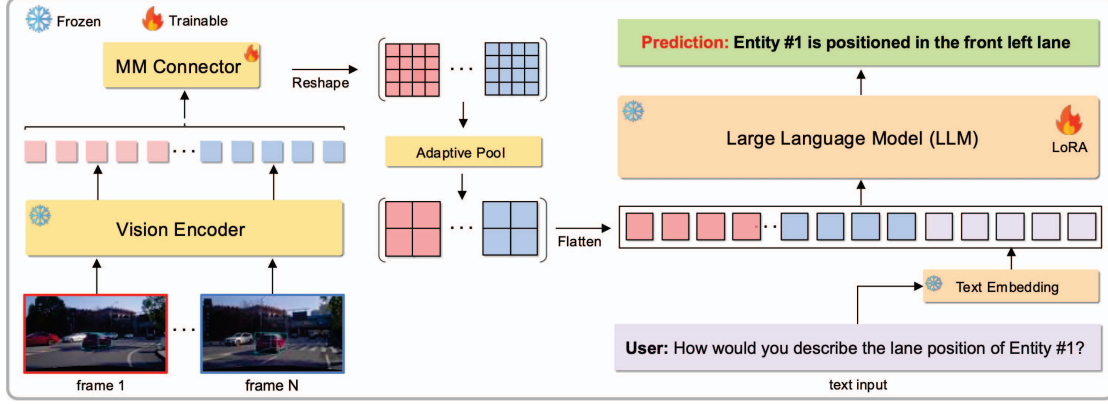


Figure 3. The overall architecture of our baseline framework.

number (i.e., 250) of samples per task. The remaining two datasets are intended for model training: the first, TB-250k, contains 250,000 samples; the second, TB-100k, includes over 100,000 samples that have been filtered to balance the number of samples per task. Table 3 summarizes the overall statistics of these datasets.

4.2. Details of the Pipeline

The Q&A pairs are generated automatically, with manual inspection following the automated process. The only exception is the ‘Other Lane Changing’ task, where we manually generate Q&A pairs due to noisy lane information at intersections. Figure 2 illustrates the pipeline used for generating Q&A pairs from these datasets.

The process unfolds as follows: The input to the pipeline is a single sample from the datasets, which could be either a static image with a set of entity attributes from KITTI/ONCE or a list of sequences with similar data from Argoverse2. The pipeline begins by extracting key information from the input, as depicted in the left panel of Fig. 2. This is followed by a processing step shown in the middle-top panel of Fig. 2, where spatial positions and facing angles relative to an anchor object are calculated. Additionally, the ‘lane to ego’ task identifies on which side the entity is located relative to the ego vehicle. For turning behaviors, we record the accumulated turning angle of each object to determine its recent motion. For lane changes, a flag is recorded if there are changes in lane_id compared to the previous step. Similarly, all sensor numerical ground truth data—such as position, dimension, and angle of all entities—are processed into attributed data, such as distance to ego and spatial position.

Finally, a rule-based process, shown in the middle-bottom panel of Fig. 2, is triggered to identify the task and generate Q&A pairs. More details are in the supp. matt.

In the next phase, a rule-based system generates QA samples from processed data attributes. This depends on the type of task, i.e., tasks aside from lane change and turn-

ing behavior can be created based on any frame, without necessarily needing an event to trigger it. Thus, they naturally have more data samples generated. After this, the rule-based QA is generated with simple short answers, such as ‘oncoming lane’ or ‘turn left.’ Then, it is augmented to be a more complex sentence using text-only information with an LLM; we used Microsoft-Phi3-medium [1].

5. Baseline Framework

We present a generic framework that serves as a strong baseline for our tasks, comprising three standard components: a vision encoder, a multi-modal connector (a two-layer MLP), and an LLM. The vision encoder extracts visual representations from input frames, the multi-modal connector projects these representations into the LLM’s embedding space, and finally the LLM generates a response based on the given question and visual embeddings. Figure 3 illustrates the architecture of our framework.

We now explain how to adaptively extract visual representations from varying numbers of frames and input them into the LLM. Given N frames of $H \times W$ having color-coded bounding boxes, the vision encoder processes each frame individually to produce N visual representations of size $[H/p \times W/p, C]$, where p is the patch size and C is the embedding dimension of the encoder. These visual representations are then projected into the LLM’s embedding space of D using the multi-modal connector, resulting in N visual embeddings of size $[H/p \times W/p, D]$.

Inputting all visual embeddings of N frames into the LLM can be computationally expensive. To address this, we sample spatially a subset of these visual embeddings per frame. Specifically, we apply adaptive average pooling to reduce each frame’s embeddings, from $[H/p \times W/p, D]$ to $[k \times h \times w, D]$, where $k \ll H/p \times W/p$. The value of k is determined as a hyperparameter. The sampled embeddings from all N frames are then reshaped and concatenated, preserving spatial and temporal order, which yields final visual embeddings of size $[N \times k, D]$ that are passed into the LLM

Table 4. Results of compared methods on TB-Bench are reported in accuracy (%), where higher indicates better performance. Random guess[†] results are considered zero. *In-context learning for single-frame tasks uses three in-context examples, while multi-frame tasks use one. Huggingface and API names are used for easy reference.

Model	TB-Bench Tasks								Avg. $\uparrow$
	RD $\uparrow$	SR $\uparrow$	OR $\uparrow$	EGO-LANE $\uparrow$	OBJ-LANE $\uparrow$	OBJ-TURN $\uparrow$	EGO-TURN $\uparrow$	EGO-TRA $\uparrow$	
Random [†]	0.0	16.7	17.1	25.0	33.3	33.3	33.3	0.0	19.8
Zero-shot									
LLaVA-1.5-7B	10.8	16.8	28.0	28.4	20.4	23.2	16.8	0.0	18.1
LLaVA-v1.6-Mistral-7B	4.0	25.6	30.8	20.4	26.0	22.4	27.2	0.0	19.6
LLaVA-NeXT-Video-7B	3.6	0.8	13.2	10.4	18.8	22.4	30.0	0.0	12.4
LLaVA-Interleave-Qwen-7B	5.6	24.8	10.8	31.6	19.2	26.8	20.4	0.0	17.4
Bunny-v1.1-4B	24.4	20.4	19.6	28.4	16.0	20.0	34.4	0.0	20.4
Bunny-v1.1-Llama-3-8B-V	7.6	16.4	30.0	26.8	18.4	21.6	20.0	1.2	17.8
InternVL2-8B	3.6	12.0	28.0	28.4	28.0	29.2	30.4	0.4	20.0
Magma-8B	1.2	24.8	11.6	34.4	36.8	46.8	42.0	2.8	25.1
Mini-InternVL2-1B-DriveLM	0.0	31.2	20.0	28.4	24.8	47.2	41.6	0.0	24.2
DriveLM-mantis-8B	0.0	34.8	23.2	30.0	57.6	50.8	48.8	0.0	30.7
Gemini-1.5-flash	21.2	16.8	22.0	34.8	48.0	23.2	27.6	4.8	24.8
GPT-4o-2024-08-06	8.4	32.0	40.8	54.4	39.6	43.2	40.4	16.0	34.4
In-context learning*									
LLaVA-Interleave-Qwen-7B	14.0	3.6	10.4	24.8	29.6	19.6	28.0	24.4	19.3
GPT-4o-2024-08-06	32.8	38.8	36.8	60.4	51.2	38.4	46.4	22.8	40.9
VLIT on TB-100k									
Ours (SigLIP-L-Qwen1.5-0.5B)	76.4	74.4	86.8	94.0	68.8	74.8	81.2	63.2	77.5
Ours (SigLIP-L-Qwen2-0.5B)	80.4	74.8	88.8	93.6	65.2	76.4	80.0	60.4	77.5
VLIT on TB-250k									
Ours (SigLIP-L-Qwen1.5-0.5B)	93.6	82.4	96.0	99.6	69.6	80.4	82.0	73.4	84.5
Ours (SigLIP-L-Qwen2-0.5B)	91.2	83.2	94.8	99.6	69.6	80.4	82.8	78.8	85.1

along with the textual embeddings.

To process text input, we tokenize the question and its ground-truth response, converting them into textual embeddings. These are then combined with the visual embeddings and input into the LLM. We train the model by minimizing cross-entropy loss on the response token predictions. During inference, only the question is used as text input.

6. Experiments

6.1. Experimental Settings

Our proposed framework is compatible with any vision encoder and LLM. We utilize pretrained SigLIP-L/14 [56] as the vision encoder and Qwen 0.5B (v1.5 or v2.0) [3, 53] as the LLM, initializing the multi-modal connector randomly. To preserve LLM capabilities and enable efficient fine-tuning, we apply LoRA [16] (rank 64). During training, we freeze the vision encoder and LLM, updating only the multi-modal connector and LoRA adapters.

For tasks requiring temporal information, the number of frames N is 8; otherwise $N = 1$. Each frame is resized to 384×384 as the input to SigLIP-L/14, with the number of sampled visual embeddings k set to 16 (i.e., $h = w = 4$).

We fine-tune our models on TB-100K or TB-250K and report accuracy on TB-Bench. Using AdamW [27] with a $2e-4$ learning rate and batch size 64, we train for 10 epochs with a cosine scheduler.

6.2. Zero-shot Evaluation for MLLMs

We report the zero-shot performance of various MLLMs on TB-Bench, including two popular proprietary models (GPT-4o, Gemini 1.5), several SOTA open-source general models, including LLaVA [26], Bunny [15], InternVL [6] and Magma [54], as well as open-source models with traffic domain adaptations trained on DriveLM [40], i.e., Mantis [18] and Mini-InternVL2 [12]. For class output questions, we use a multi-choice template listing all possible class options, while for numerical output questions, we specify the format, i.e., “Answer in xx.x meters.” See the supp. material for more details on the models and the prompt design.

6.3. Results on TB-Bench

Table 4 shows the results on TB-Bench tasks, categorized into four groups: zero-shot evaluation, in-context learning evaluation, VLIT on TB-100k, and VLIT on TB-250k.

In the zero-shot evaluation, although the proprietary

Table 5. Ablation results on (a) vision encoders, (b) number of visual embeddings per frame, and (c) number of frames.

(a) vision encoder		(b) # tokens/frame		(c) # frames	
Encoder	Acc	# tokens/fr	Acc	# frames	Acc
CLIP-L/14	72.0	4	72.7	2	72.1
SigL-B/16	74.3	16	77.5	4	73.8
SigL-L/14	77.5	36	76.2	8	77.5

Table 6. Cross-dataset generalization results of control signals prediction on BDD-X test dataset. RMSE denotes the root mean squared error, and A_τ measures the proportion of test samples with prediction errors less than τ .

Experiment	Dataset sampling ratio		Speed (m/s)					Turning angle (degree)				
	BDD-X	TB-100k	RMSE↓	$A_{0.1}$ ↑	$A_{0.5}$ ↑	$A_{1.0}$ ↑	$A_{5.0}$ ↑	RMSE↓	$A_{0.1}$ ↑	$A_{0.5}$ ↑	$A_{1.0}$ ↑	$A_{5.0}$ ↑
standard training	20	-	1.40	26.1	55.7	75.6	98.6	11.2	44.2	62.2	71.8	89.2
with co-training	20	1	1.38	26.3	57.6	76.1	98.8	11.3	44.5	63.7	73.0	89.3

models (GPT-4o and Gemini) outperform the open-source models overall, none of them excels across all traffic behavior tasks. Many open-source models underperform random guessing, while traffic domain adaptation models show significantly better performance in certain areas but still lag behind the proprietary models. The proprietary models achieve an average accuracy of less than 35%.

In in-context learning, examples significantly improve performance in specific areas, i.e., numerical outputs.

For baseline models fine-tuned on TB-100k, both with Qwen variants demonstrate strong performance across all tasks, with an average accuracy of 77.5%. Even the lowest-performing task exceeds 60% accuracy, showing a significant improvement of over almost 45% compared to GPT-4o and 57% over random chance. This underscores the effectiveness of VLIT when a high-quality dataset is available, enhancing traffic behavior understanding of MLLMs.

For baseline models fine-tuned on TB-250k, performance improves across all tasks, particularly those with increased data samples. Notably, accuracy in tasks like OBJ-LANE, OBJ-TURN, and EGO-TURN, with the same training samples to TB-100k, also benefits from additional samples in other tasks. This suggests that learning from tasks can be transferred to those with limited training data.

6.4. Ablation Study

We conduct an ablation study to identify which factors enhance performance during fine-tuning, regarding visual inputs to the models. All experiments use the same settings unless noted. The results are summarized in Table 5.

Table 5a compares different pretrained vision encoders, including CLIP-L/14 [36] and SigLIP-B/16 (processing 224×224 frames). It is seen that the SigLIP encoders outperform the CLIP encoder, with SigLIP-L/14 achieving the highest accuracy.

Table 5b presents the results of using varying numbers of sampled visual embeddings/tokens per frame, p (where $h = w = \sqrt{p}$). We observe that using 16 sampled visual tokens per frame is optimal.

Finally, we evaluate the impact of varying the number of sampled frames N ($= 2, 4, 8$), on the tasks requiring temporal information. We consistently select the first and last frames, with the remaining $N - 2$ frames sampled uniformly in between. As shown in Table 5c, increasing temporal in-

formation significantly boosts performance. For more detailed on other ablation results, see the supp. material.

6.5. Cross-Dataset Generalization

We conduct additional experiments to demonstrate performance transfer from co-training with perception-stage tasks (TB-100k) and planning tasks on the BDD-X dataset [19] to evaluate improvements in downstream tasks.

Following [12], we use Mini-InternVL2 due to its similarity in frame index referencing, inherited from its pre-training process and the BDD-X annotation. We follow a standard MLLM training regime: stage 1 focuses on feature alignment using the pre-trained Mini-InternVL2 checkpoint, while stage 2 involves instruction tuning on the mixed datasets. We apply LoRA [16] to co-train on the mixed datasets configured as in Table 6, using the same hyperparameters as in [12].

The results on Table 6 show consistent performance improvements across most metrics when co-training the MLLM with our TB-100k dataset. It indicates the benefit of our proposed dataset on improving the downstream tasks.

7. Conclusion

We have introduced TB-Bench, a comprehensive benchmark that rigorously assesses MLLM performance across eight perception tasks, providing a much-needed standard for spatiotemporal evaluation in AD. Alongside TB-Bench, we have developed the vision-language instruction tuning datasets, TB-100k and TB-250k, which significantly improve MLLM performance when used to fine-tune our baseline models, resulting in substantial gains over existing models. Additionally, our VLIT datasets offer benefits as valuable assets for mixed training datasets in other driving use cases. Our contributions represent incremental progress while laying a foundation for integrating MLLMs into AD perception, prediction, and planning stages, supporting the development of more capable and reliable autonomous driving systems. Please refer to the supp. material for further discussion on broader impact, limitations, and future work.

Acknowledgments This work was partly supported by JST [Moon-shot Research and Development], Grant Number [JPMJMS2032] and JSPS KAKENHI Grant Number 20H05952 and 23H00482.

References

- [1] Marah Abidin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Harkirat Behl, et al. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*, 2024. 6
- [2] 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. 1
- [3] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023. 7
- [4] 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. 3
- [5] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21819–21830, 2024. 3
- [6] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24185–24198, 2024. 3, 7
- [7] An-Chieh Cheng, Hongxu Yin, Yang Fu, Qiushan Guo, Ruihan Yang, Jan Kautz, Xiaolong Wang, and Sifei Liu. Spatial-rgpt: Grounded spatial reasoning in vision language model. *arXiv preprint arXiv:2406.01584*, 2024. 3, 4
- [8] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. InstructBLIP: Towards general-purpose vision-language models with instruction tuning. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. 3
- [9] 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 Conference on Empirical Methods in Natural Language Processing*, pages 2088–2098, 2019. 3
- [10] 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. 4
- [11] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Zhenyu Qiu, Wei Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, and Rongrong Ji. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*, 2023. 3, 4
- [12] Zhangwei Gao, Zhe Chen, Erfei Cui, Yiming Ren, Weiyun Wang, Jinguo Zhu, Hao Tian, Shenglong Ye, Junjun He, Xizhou Zhu, et al. Mini-internvl: A flexible-transfer pocket multimodal model with 5% parameters and 90% performance. *arXiv preprint arXiv:2410.16261*, 2024. 7, 8
- [13] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3354–3361. IEEE, 2012. 5
- [14] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research*, 32(11):1231–1237, 2013. 3
- [15] Muyang He, Yexin Liu, Boya Wu, Jianhao Yuan, Yuezhe Wang, Tiejun Huang, and Bo Zhao. Efficient multimodal learning from data-centric perspective. *arXiv preprint arXiv:2402.11530*, 2024. 3, 7
- [16] 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. In *International Conference on Learning Representations*, 2022. 7, 8
- [17] Yuichi Inoue, Yuki Yada, Kotaro Tanahashi, and Yu Yamaguchi. Nuscenesc-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. 3
- [18] Dongfu Jiang, Xuan He, Huaye Zeng, Cong Wei, Max Ku, Qian Liu, and Wenhui Chen. Mantis: Interleaved multi-image instruction tuning. *arXiv preprint arXiv:2405.01483*, 2024. 7
- [19] 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, 8
- [20] Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*, 2023. 3, 4
- [21] 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. 1, 4
- [22] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023. 3
- [23] Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, et al. Mvbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22195–22206, 2024. 3, 4

- [24] Ji Lin, Hongxu Yin, Wei Ping, Pavlo Molchanov, Mohammad Shoeybi, and Song Han. Vila: On pre-training for visual language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26689–26699, 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. 4
- [26] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024. 1, 3, 7
- [27] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 7
- [28] Arjun Majumdar, Anurag Ajay, Xiaohan Zhang, Pranav Putta, Sriram Yenamandra, Mikael Henaff, Sneha Silwal, Paul Mccvay, Oleksandr Maksymets, Sergio Arnaud, et al. Openeqa: Embodied question answering in the era of foundation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16488–16498, 2024. 3, 4
- [29] 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. 3
- [30] J Mao, M Niu, C Jiang, H Liang, J Chen, X Liang, Y Li, C Ye, W Zhang, Z Li, et al. One million scenes for autonomous driving: Once dataset. arxiv 2021. *arXiv preprint arXiv:2106.11037*. 3, 5
- [31] Ana-Maria Marcu, Long Chen, Jan Hünemann, 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. 3
- [32] Wassim G Najm, John D Smith, Mikio Yanagisawa, et al. Pre-crash scenario typology for crash avoidance research. Technical report, United States. Department of Transportation. National Highway Traffic Safety . . . , 2007. 4
- [33] 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. *arXiv preprint arXiv:2312.03661*, 2023. 3
- [34] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019. 6
- [35] Tianwen Qian, Jingjing Chen, Linhai Zhuo, Yang Jiao, and Yu-Gang Jiang. Nuscen-q: A multi-modal visual question answering benchmark for autonomous driving scenario. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 4542–4550, 2024. 3, 2
- [36] 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. 3, 8
- [37] Katrin Renz, Long Chen, Ana-Maria Marcu, Jan Hünemann, Benoit Hanotte, Alice Karnsund, Jamie Shotton, Elahe Arani, and Oleg Sinavski. Carllava: Vision language models for camera-only closed-loop driving. *arXiv preprint arXiv:2406.10165*, 2024. 1
- [38] 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. 3
- [39] Zhenwei Shao, Zhou Yu, Jun Yu, Xuecheng Ouyang, Lihao Zheng, Zhenbiao Gai, Mingyang Wang, and Jiajun Ding. Imp: Highly capable large multimodal models for mobile devices. *arXiv preprint arXiv:2405.12107*, 2024. 3
- [40] 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. In *Proceedings of the European Conference on Computer Vision*, 2024. 3, 4, 7
- [41] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, et al. Scalability in perception for autonomous driving: Waymo open dataset. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2446–2454, 2020. 3
- [42] Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023. 1
- [43] Xiaoyu Tian, Junru Gu, Bailin Li, Yicheng Liu, Chenxu Hu, Yang Wang, Kun Zhan, Peng Jia, Xianpeng Lang, and Hang Zhao. Drivelm: The convergence of autonomous driving and large vision-language models. *arXiv preprint arXiv:2402.12289*, 2024. 2, 3
- [44] Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai Charitha Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, et al. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. *arXiv preprint arXiv:2406.16860*, 2024. 2, 3, 4
- [45] Huijie Wang, Tianyu Li, Yang Li, Li Chen, Chonghao Sima, Zhenbo Liu, Bangjun Wang, Peijin Jia, Yuting Wang, Shengyin Jiang, et al. Openlane-v2: A topology reasoning benchmark for unified 3d hd mapping. *Advances in Neural Information Processing Systems*, 36, 2024. 3
- [46] Benjamin Wilson, William Qi, Tanmay Agarwal, John Lambert, Jagjeet Singh, Siddhesh Khandelwal, Bowen Pan, Ratnesh Kumar, Andrew Hartnett, Jhony Kaesemodel Pontes, Deva Ramanan, Peter Carr, and James Hays. Argoverse 2: Next generation datasets for self-driving perception and forecasting. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks (NeurIPS Datasets and Benchmarks 2021)*, 2021. 3, 5

- [47] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. Huggingface’s transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*, 2019. 6
- [48] Dongming Wu, Wencheng Han, Tiancai Wang, Xingping Dong, Xiangyu Zhang, and Jianbing Shen. Referring multi-object tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14633–14642, 2023. 3
- [49] Dongming Wu, Wencheng Han, Tiancai Wang, Yingfei Liu, Xiangyu Zhang, and Jianbing Shen. Language prompt for autonomous driving. *arXiv preprint arXiv:2309.04379*, 2023. 3
- [50] Penghao Wu and Saining Xie. V?: Guided visual search as a core mechanism in multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13084–13094, 2024. 2
- [51] Li Xu, He Huang, and Jun Liu. Sutr-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. 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9523–9532, 2020. 3
- [53] An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2024. 7
- [54] Jianwei Yang, Reuben Tan, Qianhui Wu, Ruijie Zheng, Baolin Peng, Yongyuan Liang, Yu Gu, Mu Cai, Seonghyeon Ye, Joel Jang, et al. Magma: A foundation model for multimodal ai agents. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. 7
- [55] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9556–9567, 2024. 3, 4
- [56] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11975–11986, 2023. 7
- [57] Peiyuan Zhang, Kaichen Zhang, Bo Li, Guangtao Zeng, Jingkan Yang, Yuanhan Zhang, Ziyue Wang, Haoran Tan, Chunyuan Li, and Ziwei Liu. Long context transfer from language to vision. *arXiv preprint arXiv:2406.16852*, 2024. 1, 4
- [58] Weize Zhang, Mohammed Elmaghiubi, Kasra Rezaee, Behzad Khamidehi, Hamidreza Mirkhani, Fazel Arasteh, Chunlin Li, Muhammad Ahsan Kaleem, Eduardo R Corral-Soto, Dhruv Sharma, et al. Analysis of a modular autonomous driving architecture: The top submission to carla leaderboard 2.0 challenge. *arXiv preprint arXiv:2405.01394*, 2024. 1
- [59] YiFan Zhang, Huanyu Zhang, Haochen Tian, Chaoyou Fu, Shuangqing Zhang, Junfei Wu, Feng Li, Kun Wang, Qingsong Wen, Zhang Zhang, Liang Wang, and Rong Jin. MME-realworld: Could your multimodal LLM challenge high-resolution real-world scenarios that are difficult for humans? In *International Conference on Learning Representations*, 2025. 3, 4
- [60] 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