

DriveLM: Driving with Graph Visual Question Answering

Chonghao Sima^{4,1*} Katrin Renz^{2,3*} Kashyap Chitta^{2,3} Li Chen^{4,1}
 Hanxue Zhang¹ Chengen Xie¹ Jens Beißwenger^{2,3} Ping Luo⁴
 Andreas Geiger^{2,3†} Hongyang Li^{1†}

¹ OpenDriveLab, Shanghai AI Lab, China

² University of Tübingen, Germany

³ Tübingen AI Center, Germany

⁴ University of Hong Kong, China

simachonghao@pjlab.org.cn katrin.renz@uni-tuebingen.de

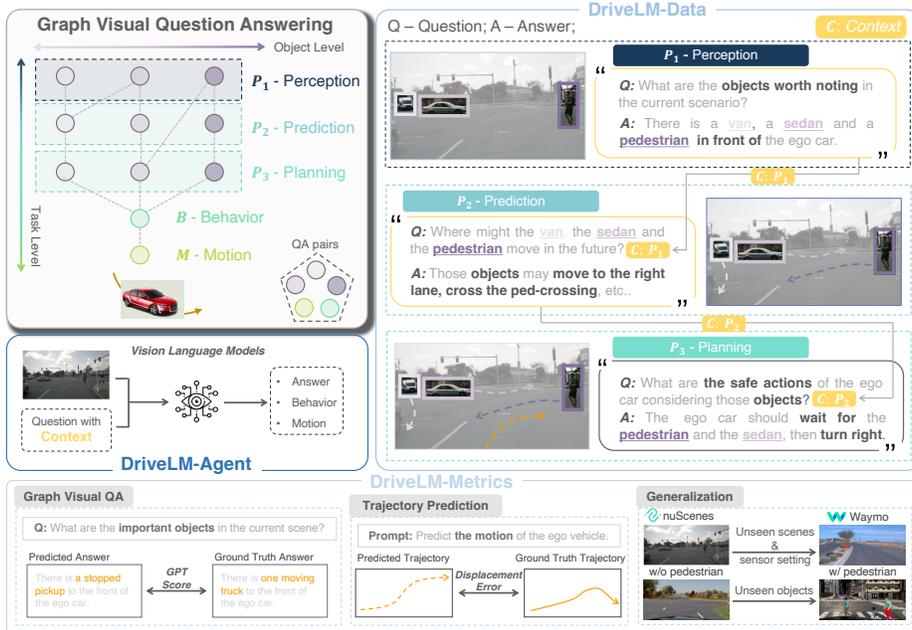


Fig. 1: We present **DriveLM**: A new task, dataset, metrics, and baseline for end-to-end autonomous driving. Inspired by [10], DriveLM considers **Graph Visual Question Answering (GVQA)**, where QA pairs are interconnected via logical dependencies at the object-level, *i.e.*, interactions between object pairs, and the task-level, *e.g.*, perception → prediction → planning → behavior (discretized action described in natural language) → motion (continuous trajectory). We propose **DriveLM-Data** for training **DriveLM-Agent**, a baseline for GVQA. We validate its effectiveness using the **DriveLM-Metrics** on challenging settings requiring zero-shot generalization.

Abstract. We study how vision-language models (VLMs) trained on web-scale data can be integrated into end-to-end driving systems to

*Equal contribution. †Project lead.

boost generalization and enable interactivity with human users. While recent approaches adapt VLMs to driving via single-round visual question answering (VQA), human drivers reason about decisions in multiple steps. Starting from the localization of key objects, humans estimate object interactions before taking actions. The key insight is that with our proposed task, Graph VQA, where we model graph-structured reasoning through perception, prediction and planning question-answer pairs, we obtain a suitable proxy task to mimic the human reasoning process. We instantiate datasets (DriveLM-Data) built upon nuScenes and CARLA, and propose a VLM-based baseline approach (DriveLM-Agent) for jointly performing Graph VQA and end-to-end driving. The experiments demonstrate that Graph VQA provides a simple, principled framework for reasoning about a driving scene, and DriveLM-Data provides a challenging benchmark for this task. Our DriveLM-Agent baseline performs end-to-end autonomous driving competitively in comparison to state-of-the-art driving-specific architectures. Notably, its benefits are pronounced when it is evaluated zero-shot on unseen sensor configurations. Our question-wise ablation study shows that the performance gain comes from the rich annotation of prediction and planning QA pairs in the graph structure. All data, models and an official evaluation server are available at <https://github.com/OpenDriveLab/DriveLM>.

Keywords: Vision Language Model · End-to-end Autonomous Driving

1 Introduction

Current Autonomous Driving (AD) stacks are still lacking crucial capabilities [10, 15]. One key requirement is generalization, which involves the ability to handle unseen scenarios or unfamiliar sensor configurations. A secondary requirement pertains to the interaction of these models with humans, highlighted for example by EU regulations that mandate explainability in deployment [3]. Furthermore, unlike today’s AD models, humans do not navigate based on geometrically precise bird’s-eye view (BEV) representations [17, 32, 46]. Instead, humans implicitly perform object-centric perception, prediction, and planning (which we refer to as P_{1-3}): a rough identification and localization of key objects, followed by reasoning about their possible movement and aggregation of this information into a driving action [60, 76].

Simultaneously, another field has been forging ahead: Vision-Language Models (VLMs) [47, 54, 90, 106]. These models have several strengths. First, they hold a base understanding of the world from internet-scale data that could potentially facilitate generalization for planning in AD. In fact, this sort of generalization has already been achieved by VLMs for simpler robotics tasks [23, 108]. Second, the use of language representations as an input and output offers a platform for human-friendly interaction with these models, unlike bounding boxes or trajectories that are more common to current methods [19, 31, 49, 70]. Finally, VLMs are able to make decisions in multiple steps linked by logical rea-

soning [5, 21, 93, 95, 103, 108]. Importantly, even though they reason in multiple separate steps, VLMs are end-to-end differentiable architectures, a characteristic that is highly desirable for autonomous driving [10].

Recent work towards enabling the application of VLMs to AD systems falls into two categories: scene-level or single object-level Visual Question Answering (VQA). Scene-level VQA refers to the task of describing the driving behavior by one or two supporting reasons, *e.g.*, “The car is moving into the right lane because it is safe to do so.” [41, 42]. Single object-level VQA formulates the understanding of the ego vehicle’s response to a single object by a chain of QAs in the form of “what-which-where-how-why”, *e.g.*, “The ego vehicle stops because there is a pedestrian in a white shirt crossing the intersection in front of the ego vehicle and it does not want to crash into the pedestrian.” [56, 67, 71]. Unfortunately, neither of these paradigms provides a suitable proxy task to mimic the P_{1-3} reasoning process in humans, who consider multiple objects and reason about each in multiple steps. Therefore, in this paper, we propose a new task, along with corresponding datasets and a baseline model architecture (Fig. 1).

Task. Graph Visual Question Answering (GVQA) involves formulating P_{1-3} reasoning as a series of question-answer pairs (QAs) in a directed graph. Its key difference to the aforementioned VQA tasks for AD is the availability of logical dependencies between QAs which can be used to guide the answering process. GVQA also encompasses questions regarding behavior and motion planning, with dedicated metrics (details in Section 2).

Datasets. DriveLM-nuScenes and **DriveLM-CARLA** consist of annotated QAs, arranged in a graph, linking images with driving behavior through logical reasoning. In comparison to existing benchmarks, they provide significantly more text annotations per frame (Table 1 and Fig. 2). Furthermore, as an integral component of DriveLM-CARLA, we build PDM-Lite [4], the first working rule-based expert algorithm for CARLA Leaderboard 2.0. We pair these training datasets with challenging test data for evaluating zero-shot generalization.

Baseline. DriveLM-Agent employs a trajectory tokenizer that can be applied to any general VLM [47, 54, 65, 106], coupled with a graph prompting scheme that models logical dependencies as context inputs for VLMs. The result is a simple methodology to effectively repurpose VLMs for end-to-end AD (Section 3).

Our experiments provide encouraging results. We find that GVQA on DriveLM is a challenging task, where current methods obtain moderate scores and better modeling of logical dependencies is likely necessary to achieve strong QA performance. Even so, DriveLM-Agent already performs competitively to state-of-the-art driving-specific models [32] when tested in the open-loop planning setting, despite its task-agnostic architecture. Furthermore, employing a graph structure improves zero-shot generalization, enabling DriveLM-Agent to correctly handle unseen deployment on Waymo data [77] after training only on nuScenes [6] data. Finally, with a question-wise analysis we find that the QA pairs from the prediction and planning stages help the final driving decision most. From these results, we believe that improving GVQA holds great potential towards building autonomous driving agents with strong generalization.

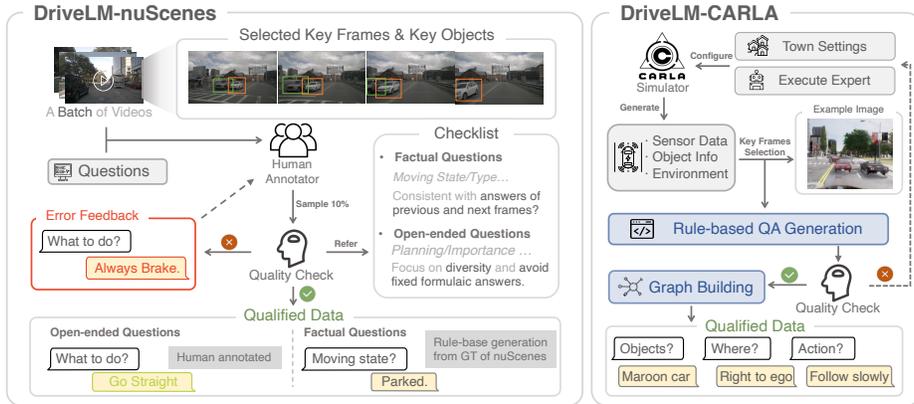


Fig. 2: Annotation Pipeline: In DriveLM-nuScenes, we adopt a semi-rule-based QA labeling pipeline, where both the ground truth annotation in nuScenes/OpenLane-V2 and feedback from human annotators are used. A critical part of our pipeline is the multi-round quality check, which guarantees high data quality at reasonable costs. In DriveLM-CARLA, we meet the same standards while exploiting a fully rule-based QA labeling pipeline instead, using a new expert algorithm called PDM-Lite.

2 DriveLM: Task, Data, Metrics

Human drivers usually decompose their decision-making process into distinct stages that follow a logical progression which encompasses the identification and localization of key objects, their possible future action and interaction, and ego planning based on all this information [27, 55]. This inspires us to propose the GVQA as the critical ingredient of DriveLM, which serves as a suitable proxy task to mimic the human reasoning process. Within this section, we illustrate the formulation of the GVQA task (Section 2.1), introduce DriveLM-Data (Section 2.2) to exemplify the instantiation of GVQA using prominent driving datasets, and overview the DriveLM-Metrics used for evaluation (Section 2.3). To encourage further research in this direction, an official evaluation server (with a public leaderboard) will be set up to benchmark different methods and discover more insights about combining language models with autonomous driving.

2.1 DriveLM-Task: GVQA

We organize all the question-answer pairs (QAs) for an image frame into a graph. We use the terminology “graph-structured” to refer to a directed acyclic graph (DAG), e.g., the current question can get context from (multiple) parent and grandparent nodes. The graph $G = (V, E)$ contains a set of vertices V , where each vertex represents a QA pair $v = (q, a)$ associated with one or more key objects in the scenario. The key difference between GVQA and ordinary VQA is that the QAs in GVQA have logical dependencies, which we formulate as the edges between the vertices. $E \subseteq V \times V$, is a set of directed edges, where each edge $e = (v_p, v_c)$ connects the parent QA and the child QA. We formulate the

Table 1: Comparison of DriveLM-nuScenes & -CARLA with Existing Datasets. DriveLM-Data significant advances annotation quantity, comprehensiveness (covering **perception, prediction and planning**), and logic (chain to **graph**). † full dataset, ‡ keyframe dataset, * semi-rule-based labeling (w/ human annotators), ** fully-rule-based (no human annotators). - means publicly unavailable.

Dataset	Source Dataset	# Frames	Avg. QA per Frame	Perception	Prediction	Planning	Logic
nuScenes-QA [67]	nuScenes	34,149	13.5	460k**	✗	✗	None
nuPrompt [98]	nuScenes	34,149	1.0	35k*	✗	✗	None
HAD [41]	HDD	25,549	1.8	25k	✗	20k	None
BDD-X [42]	BDD	26,228	1	26k	✗	✗	None
LingoQA [59]	LingoQA	28,000	15.3	-	-	-	None
DRAMA [56]	DRAMA	17,785	5.8	85k	✗	17k	Chain
Rank2Tell [71]	Rank2Tell	5,800	-	-	✗	-	Chain
DriveLM-nuScenes	nuScenes	4,871	91.4	144k*	153k	146k	Graph
DriveLM-CARLA†	CARLA	64,285	24.4	697k**	311k**	558k**	Graph
DriveLM-CARLA‡	CARLA	5,721	24.8	63k**	28k**	51k**	Graph

edge set E by incorporating two dimensions: object-level and task-level edges. At the object level, we construct the logical edges $e \in E$ to represent the impact of interactions between different objects. For example, the planning QA node for the sedan is influenced by the perception QA node of the pedestrian in the illustration from Fig. 1 (center). At the task-level, we establish the logical edges $e \in E$ to capture the logical chain of different reasoning stages:

- **Perception** (P_1): identification, description, and localization of key objects in the current scene.
- **Prediction** (P_2): estimation of possible action/interaction of key objects based on perception results.
- **Planning** (P_3): possible safe actions of the ego vehicle.
- **Behavior** (B): classification of driving decision.
- **Motion** (M): waypoints of ego vehicle future trajectory.

The concepts of perception, prediction, and planning (P_{1-3}) are similar to those in end-to-end AD [10], while the concepts of motion and behavior are based on the ego vehicle future trajectory. Specifically, we define the motion M as the ego vehicle future trajectory, which is a set of N points with coordinates (x, y) in bird’s-eye view (BEV), denoted as $M = \{(x_0, y_0), (x_1, y_1), \dots, (x_N, y_N)\}$. Each point is the offset between the future position and the current position by a fixed time interval. Then, the distance for x, y at each time interval is computed as:

$$\{x, y\}_{\text{dist}} = \{(\delta_{x,1}, \delta_{y,1}), \dots, (\delta_{x,N}, \delta_{y,N})\}, \quad (1)$$

where $\delta_{x,i} = x_i - x_{i-1}$ and $\delta_{y,i} = y_i - y_{i-1}$, for $i = 1, 2, \dots, N$. The goal of the behavior representation is to serve as an interface from P_{1-3} to M . To obtain a behavior representation, we map the mean of x_{dist} and y_{dist} to one of the predefined bins, where each bin corresponds to a category in either speed or steering. These are denoted as B_{sp} and B_{st} respectively. In this work, we consider 5 bins:

$$\begin{aligned}
B_{sp} &\in \{\text{fast}_2, \text{fast}_1, \text{moderate}, \text{slow}_1, \text{slow}_2\}, \\
B_{st} &\in \{\text{left}_2, \text{left}_1, \text{straight}, \text{right}_1, \text{right}_2\},
\end{aligned}$$

where the number in the subscript indicates the intensity. The combination of the speed and steering categories for a trajectory form its behavior category as $B = (B_{sp}, B_{st})$. While we use a simple definition of B as a starting point for research on driving with VLMs, we note that our formulation supports the incorporation of more abstract behaviors such as a lane changes or overtaking.

2.2 DriveLM-Data

In order to provide comprehensive and accurate QAs with the graph structure defined in Section 2.1, we introduce DriveLM-nuScenes and DriveLM-CARLA. Since there are significant disparities between nuScenes and CARLA, the collection methods and statistics of these datasets differ.

DriveLM-nuScenes. We divide the annotation process into three steps: selecting key frames from video clips, choosing key objects within these key frames, and subsequently annotating the frame-level P_{1-3} QAs for these key objects. A portion of the Perception QAs are generated from the nuScenes [6] and OpenLane-V2 [87] ground truth, while the remaining QAs are manually annotated. The question templates for the manual annotations were designed by 5 domain experts accounting for how humans make driving decisions. Annotators are prompted with all question templates for each frame, encouraged to answer all questions, but provided a skip option to account for possible incompatibility. As we manually annotate the vast majority of data in DriveLM-nuScenes, quality is particularly crucial for this portion. When annotating, we conduct multiple rounds of rigorous quality checks. In each round, we categorize the data into different batches and inspect ten percent of the data in each batch. If the qualification rate of manually annotated data in this ten percent does not meet expectations, we request the annotators to re-label all data in the batch. In Fig. 2, we showcase an example of the QA annotation pipeline, where all questions undergo quality checks according to our standards. As a result, DriveLM-nuScenes stands out from previously proposed datasets with its larger scale, greater comprehensiveness, and more complex structure (See Table 1). These QAs cover various aspects of the driving process, ranging from perception and prediction to planning, providing a comprehensive understanding of autonomous driving scenarios (details in the supplementary material).

DriveLM-CARLA Expert. We collect data using CARLA 0.9.14 in the Leaderboard 2.0 framework [22]. Leaderboard 2.0 contains a large set of new driving scenarios compared to its predecessor, Leaderboard 1.0. However, as of now, there is no existing method to collect training data at scale in Leaderboard 2.0. For example, the privileged rule-based expert used by TransFuser++ [36], a state-of-the-art method in Leaderboard 1.0, obtains a Driving Score (DS) of only 2% on the 8+ kilometer long official validation routes of Leaderboard 2.0. We build a new expert algorithm, PDM-Lite, that can handle the new challenges

in Leaderboard 2.0. Our approach is similar to PDM-Closed [19], a rule-based planner for nuPlan [7]. PDM-Lite uses the Intelligent Driver Model (IDM) [84] to obtain a target speed based on the leading vehicle, pedestrian, stop sign or traffic light. Unlike the 16 proposals from which one is selected via a complex cost function in PDM-Closed, we use only two proposals and a simpler cost function based on the TransFuser++ expert [36], giving a light-weight planner suitable for scalable QA generation. PDM-Lite obtains an improved DS of 44% on the official CARLA validation routes. More details can be found in the supplementary.

DriveLM-CARLA QA Generation. For collecting the DriveLM-CARLA dataset, we set up a series of routes in urban, residential, and rural areas and execute PDM-Lite on these routes. During this process, we collect the necessary sensor data, sample keyframes, generate relevant QAs based on privileged information about objects and the scene, and organize the logical relationships to connect this series of QAs into a graph. We generate data and labels at 4 FPS and extract keyframes based on changes in the decision of the expert (e.g. when the expert changes from acceleration to braking). The rule-based annotation pipeline is illustrated in Fig. 2. During data collection, we extract privileged information from the simulator about the status of the static and dynamic objects in the scene as well as the triggered rules of the expert. We use all 38 scenarios except for InterurbanAdvancedActorFlow, MergerIntoSlowTraffic, and VehicleTurningRoute to create questions and answers with hand-crafted sentence templates based on the information we extract from the simulator. The exact questions with their graph structure can be found in the supplementary. Our annotation process has the advantage of straightforward scalability since we only need to define route and scenario settings in CARLA and the subsequent steps can be executed automatically. Including 1.6M QAs (with a straightforward scaling recipe), our DriveLM-CARLA stands out as the largest driving-language benchmark in terms of total textual content among existing benchmarks as shown in Table 1.

2.3 DriveLM-Metrics

To evaluate GVQA, the DriveLM-Metrics consist of three components for evaluating motion M , behavior B , and P_{1-3} . For measuring the performance of the motion stage, we use standard metrics from the nuScenes and Waymo benchmarks: average and final displacement error, (**ADE**, **FDE**), and the **collision rate** on the predicted trajectory, following UniAD [32]. We evaluate behavior predictions by the **classification accuracy**, along with a breakdown of the overall accuracy into its steering and speed components. Finally, we measure the P_{1-3} performance using two metrics. **SPICE** [2] is a prevailing metric used in VQA and image captioning, which calculates the structure similarity of predicted texts with ground truth while ignoring the semantic meanings. Simultaneously, we employ **GPT Score** to measure the semantic alignment of answers and complement the SPICE metric. Specifically, the question, the ground truth answer, the predicted answer, and a prompt asking for a numerical score of the answer

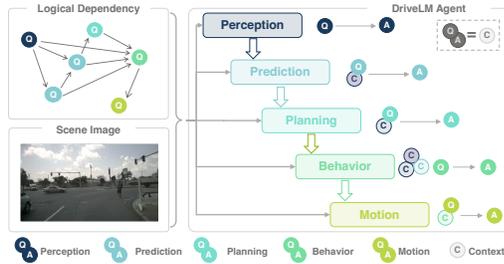


Fig. 3: DriveLM-Agent Pipeline. Given the scene image, a VLM performs prompting with context to model the logical dependency among the five QA stages. Context is built using preceding QAs, and can have one or more sources.

are sent to ChatGPT-3.5 [61, 62]. We parse the text returned to get the score, where a higher score indicates better semantic accuracy.

3 DriveLM-Agent: A GVQA Baseline

In this section, we present DriveLM-Agent, a baseline approach for the GVQA task detailed in Section 2. DriveLM-Agent is built upon a general vision-language model and can therefore exploit underlying knowledge gained during pre-training. Our overall goal involves translating an image into the desired ego vehicle motion (M) through the different stages of VQA (P_1, P_2, P_3, B). For this, we choose BLIP-2 [47] as our base VLM due to its simplicity in architecture and flexibility in fine-tuning, but our approach can be applied agnostically to other VLMs.

As shown in Fig. 3, DriveLM-Agent can be decomposed into several stages: (1) P_{1-3} , *i.e.*, *perception*, *prediction*, *planning*, serve as the foundational layers to understand the scene and reason about its structure. (2) The *behavior* stage aggregates crucial information from the P_{1-3} into a description of the desired driving action in language space. (3) Finally, the *motion* stage translates the behavior into an executable driving trajectory. To implement the logical dependency between each linked QA, we propose to use context between connected nodes in the GVQA graph. We expand on this idea in the following.

3.1 Prompting with Context

Directly translating images to motion as in [16, 66] is extremely challenging. Motivated by the tendency of humans to perform a multi-step reasoning process, we propose to use a similar strategy for VLM-based driving. By doing so, we facilitate the retrieval of knowledge stored in LLMs and improve explainability.

More precisely, the model is designed to use answers from the previous steps in the reasoning process as the context for the following questions. For each edge $e = (v_p, v_c) \in E$, we append the QA from the parent node v_p to the question of the current node v_c with a prefix “*Context:*”. The context can also contain QAs from multiple preceding nodes in which case we concatenate all QAs to one context sequence. It is worth noting that the context is only one possible implementation to formulate logical dependency in GVQA, which we select due to its simplicity. With this scheme, we pass forward relevant information based on the logical dependencies established by the graph.

Note that the size and structure of the graph during inference is a design choice of the algorithm, which can be adapted based on the task or available compute budget. We use this property to train on all available QAs, but perform inference on specific subgraphs, where the questions are sampled using heuristics. For more details, please refer to the supplementary material.

Context Aggregation through Behavior. Driving encompasses a wide array of potential situations that require an appropriate response. However, despite the diversity of these circumstances, it is interesting to note that almost all events involve decisions that can be discretized into a set of behaviors. For example, applying the brakes appropriately may address various situations such as a red light signal, a stop sign, or the presence of an object ahead of the vehicle. The focus of our behavior stage is to generate such a behavior: a statement in natural language that articulates the vehicle’s intended movement. As described in Section 2.1, this description effectively serves as a reflective step wherein the model summarizes all crucial information from the graph. Thus, we propose to use all possible sources of context for predicting behavior, *i.e.*, all the QAs in P_{1-3} . We empirically observe that this design is crucial for driving with VLMs.

3.2 Trajectory Tokenization for Motion

Since it is non-trivial to output fine-grained numerical results using general VLMs, RT-2 [108] handles robotic actions based on a specialized trajectory tokenization module. We use this approach to enable DriveLM-Agent to take as input the image and behavior description and output trajectories. Specifically, we divide the coordinates of waypoints into 256 bins empirically based on statistics of train set trajectories. We re-define tokens in the BLIP-2 language tokenizer, establishing tokens for each bin, and fine-tune the VLM on the re-defined vocabulary. For simplicity, we use the same VLM architecture (BLIP-2) to perform this task, but with independent LoRA weights and trained on a dataset consisting of only the QAs for this motion stage. Thus, it is possible to perform this functionality using a lightweight LLM [68] or driving-specific architecture that accepts a command as an input [31, 99].

4 Experiments

In this section, we present our experimental results that aim to address the following research questions: (1) How can VLMs be effectively repurposed for end-to-end autonomous driving? (2) Can VLMs for driving generalize when evaluated with unseen sensor setups? (3) What is the effect of each question in perception, prediction and planning on the final behavior decision? (4) How well do VLMs perform perception, prediction, and planning via GVQA? Experiments about more VLMs on DriveLM-nuScenes and generalization to unseen objects on DriveLM-CARLA are included in the supplementary material.

Setup. We now briefly overview the key implementation details for the two settings used in our experiments (additional details are provided in the supple-

mentary material). All fine-tuning is implemented with LoRA [30]. On DriveLM-nuScenes, we finetune BLIP-2 on the `train` split for 10 epochs. We use a batch size of 2 for each GPU, and the entire training process spans approximately 7 hours with 8 V100 GPUs. We train BLIP-2 on the keyframes of the `train` split of DriveLM-CARLA for 6 epochs. This takes 6 hours on 4 A100 GPUs.

4.1 VLMs for End-to-End Driving

In our first experiment, we aim to assess the ability of VLMs to perform open-loop planning on DriveLM-nuScenes. In particular, we investigate the impact of the context provided to the behavior and motion stages. Given sensor data (and in the case of VLM methods, a text input), the model is required to predict the ego-vehicle future trajectory in the form of waypoints. Though open-loop planning suffers from a mismatched data distribution [50], we try to alleviate this by evaluating on only the key frames annotated in DriveLM-nuScenes.

Baselines. As a reference for the task’s difficulty, we provide a simple **Command Mean** baseline. Each frame in nuScenes is associated with one of 3 commands, ‘turn left’, ‘turn right’, or ‘go straight’. We output the mean of all trajectories in the training set whose command matches the current test frame command. Further, we compare our approach to the current state-of-the-art on nuScenes, UniAD [32]. Besides its original setting which requires video inputs, we train a single-frame version (‘**UniAD-Single**’) for a fair comparison. Finally, **BLIP-RT-2** denotes BLIP-2 [47] fine-tuned on DriveLM-Data with the trajectory tokenization scheme described in Section 3.2 for only the motion task. This acts as an indicator for the performance when using an identical network architecture as DriveLM-Agent, but no context inputs or VQA training data.

DriveLM-Agent. We consider 3 variants of DriveLM-Agent incorporating our proposed changes in steps: (1) a 2-stage version that predicts behavior and then motion (as described in Section 2.1), but without any P_{1-3} context for behavior prediction (‘None’); (2) a ‘Chain’ version that builds the P_{1-3} graph, but only passes the final node (P_3) to the behavior stage; (3) the full model (‘Graph’) that uses all QAs from P_{1-3} as context for B .

Results. We show results for the above methods in Table 2. Among the baselines, BLIP-RT-2 is unable to match UniAD-Single (though both perform well relative to Command Mean). This shows that the single-stage approach without any reasoning is unable to compete with the prior state-of-the-art on nuScenes. However, the proposed DriveLM-Agent, which predicts behavior as an intermediate step for motion, provides a significant boost in performance, surpassing UniAD-Single. This indicates that with the appropriate prompting, VLMs can be surprisingly competitive for end-to-end driving. Interestingly, in the experimental settings that do not involve generalization, the Chain and Graph versions of DriveLM-Agent do not provide any further advantage over no context. Further, single-frame VLMs fall short compared to the privileged video-based UniAD model, indicating that VLMs with video inputs may be necessary for this task. We provide video-input VLM result in the supplementary material.

Table 2: Open-loop Planning on DriveLM-nuScenes and Zero-shot Generalization across Sensor Configurations on Waymo. Under nuScenes, using Behavior (B) as context for Motion (M) enables end-to-end driving with VLMs on par with UniAD-Single, a state-of-the-art driving-specific architecture. Under Waymo, we randomly sampled 1k frames from the Waymo val set after training on DriveLM-nuScenes. DriveLM-Agent outperforms UniAD-Single and benefits from graph context.

Method	Behavior Motion		nuScenes				Waymo					
	(B)	(M)	Behavior (B)		Motion (M)		Behavior (B)		Motion (M)			
	Context	Context	Acc. \uparrow	Speed \uparrow	Steer \uparrow	ADE \downarrow	Col. \downarrow	Acc. \uparrow	Speed \uparrow	Steer \uparrow	ADE \downarrow	Col. \downarrow
Command Mean	-	-	-	-	-	4.57	5.72	-	-	-	7.98	11.41
UniAD-Single	-	-	-	-	-	1.80	2.62	-	-	-	4.16	9.31
BLIP-RT-2	-	-	-	-	-	2.63	2.77	-	-	-	2.78	6.47
DriveLM-Agent	None	B	61.45	72.20	84.73	1.39	1.67	35.70	43.90	65.20	2.76	6.59
	Chain	B	50.43	60.32	75.34	2.07	2.08	34.62	41.28	64.55	2.85	6.89
	Graph	B	57.49	69.89	80.63	1.74	1.89	39.73	54.29	70.35	2.63	6.17
UniAD [32]	-	-	-	-	-	0.80	0.17	-	-	-	-	-

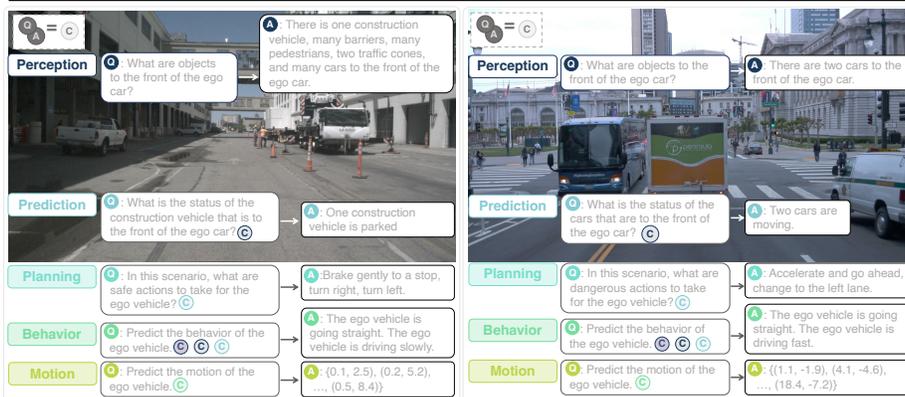


Fig. 4: Qualitative Results of DriveLM-Agent. (Left) DriveLM-nuScenes val frame, (Right) Waymo val frame. We show the questions (Q), context (C), and predicted answers (A). DriveLM-Agent’s outputs are easy to interpret for human users.

4.2 Generalization Across Sensor Configurations

As a more challenging setting for evaluating the models from Section 4.1, we now apply them without further training to a new domain: the Waymo dataset [77]. Waymo does not include rear cameras’ images, so we drop this input for UniAD-Single. VLM methods only use the front view and do not require any adaptation.

Results. As shown in Table 2, UniAD-Single does not cope well with the new sensor configuration, and drops below BLIP-RT-2 in performance. The multi-stage approach of DriveLM-Agent provides further improvements. In particular, the accuracy of speed predictions rises from 43.90 with no context to 54.29 with the full graph. On the other hand, the chain approach does not provide sufficient useful information, with a speed accuracy of only 41.28.

We present qualitative results for DriveLM-Agent on nuScenes and Waymo in Fig. 4. The model generally provides intuitive answers, with a few exceptions (*e.g.*, planning on DriveLM-nuScenes, perception on Waymo). This shows the

utility of GVQA towards interactive driving systems. Further, on Waymo, we see meaningful prediction and planning answers despite the imperfect perception. For more visualizations, please see the supplementary material.

4.3 Question-wise Analysis in DriveLM-nuScenes

Next, we analyze the effect of each type of QA pair on the behavior performance by adding them as context. First, a set of representative QA pairs are selected in each stage. We then train BLIP-2 on different combinations of sets of QA pairs, and those QA pairs are added as the context for the behavior question (Table 3).

Representative Questions at each stage. We select different sets of representative QA pairs in each stage as follows:

- P_{1-1} : What are the important objects in the current scene?
- P_{1-2} : What is the moving status of object X ?
- P_{1-3} : What is the visual description of object X ?
- P_{2-1} : What is the future state of object X ?
- P_{2-2} : Would object X be in the moving direction of the ego vehicle?
- P_{2-3} : What object should the ego vehicle notice first / second / third when the ego vehicle is getting to the next possible location?
- P_{3-1} : What actions could the ego vehicle take based on the observation of object X ?
- P_{3-2} : What actions taken by the ego vehicle can lead to a collision with object X ?
- P_{3-3} : In this scenario, what are safe actions to take for the ego vehicle?

The reasons for such a selection are **1)** statistically they make up about 60% of the total QA pairs in the three stages. **2)** subjectively they represent the most information needed for human to drive. A coarser stage-wise analysis is provided in the supplementary material for further investigation.

Results. Our results are shown in Table 3. We observe that training with QA pairs from prediction and planning stage (**ID** 4-9) improves the performance from training with perception QA pairs only (**ID** 1-3). Adding QA pairs from the planning stage (**ID** 7-9) does not significantly boost the performance compared to their previous stages (**ID** 4-6). The reason could be that other vehicles’ future status contains all the necessary information to make the behavior decision.

4.4 Performance for P_{1-3} via GVQA

In our final experiment, we establish baseline results for the P_{1-3} stages of GVQA, studying the impact of context. We use two VLMs, the off-the-shelf BLIP-2 [47] (not fine-tuned on DriveLM), and the proposed DriveLM-Agent.

Baselines. We consider the lower bound of no context (‘None’), which corresponds to training and evaluation with the same setting as standard VQA (image and question in, answer out). As an upper bound for each architecture, we perform GVQA but input the ground truth (‘GT’) context to the model at test time instead of its own prior predictions.

Results. Our results are summarized in Table 4. It can be observed that DriveLM-nuScenes is more challenging for both models, as indicated by the lower scores on

Table 3: Question-wise analysis in DriveLM-nuScenes. Questions of P_{x-y} listed in Section 4.3. Using Prediction and Planning stages of QA pairs as context for Behavior question improves the performance from using Perception only. However, the performance of different questions in the identical stage differentiate slightly.

ID	Perception			Prediction			Planning			Behavior		
	P_{1-1}	P_{1-2}	P_{1-3}	P_{2-1}	P_{2-2}	P_{2-3}	P_{3-1}	P_{3-2}	P_{3-3}	Acc. \uparrow	Speed \uparrow	Steer \uparrow
1	✓	-	-	-	-	-	-	-	-	54.69	66.83	75.22
2	✓	✓	-	-	-	-	-	-	-	55.32	67.33	74.34
3	✓	✓	✓	-	-	-	-	-	-	53.94	65.33	75.72
4	✓	✓	✓	✓	-	-	-	-	-	58.82	71.83	80.98
5	✓	✓	✓	✓	✓	-	-	-	-	57.07	71.96	78.97
6	✓	✓	✓	✓	✓	✓	-	-	-	57.70	72.22	79.22
7	✓	✓	✓	✓	✓	✓	✓	-	-	58.95	72.59	80.23
8	✓	✓	✓	✓	✓	✓	✓	✓	-	57.95	72.97	79.97
9	✓	✓	✓	✓	✓	✓	✓	✓	✓	57.49	69.89	80.63

Table 4: Baseline P_{1-3} Results. DriveLM-Agent and the zero-shot BLIP-2 benefit from a step-wise reasoning procedure given by our graph structure.

Context	DriveLM-nuScenes (P_{1-3})				DriveLM-CARLA (P_{1-3})			
	BLIP-2 [47]		DriveLM-Agent		BLIP-2 [47]		DriveLM-Agent	
	SPICE \uparrow	GPT \uparrow	SPICE \uparrow	GPT \uparrow	SPICE \uparrow	GPT \uparrow	SPICE \uparrow	GPT \uparrow
None	4.34	42.97	42.56	71.39	10.46	46.37	72.71	79.67
Graph	7.71	45.21	49.54	72.51	10.30	55.03	75.26	81.78
<i>GT</i>	<i>8.19</i>	<i>41.10</i>	<i>50.29</i>	<i>72.94</i>	<i>16.18</i>	<i>57.98</i>	<i>79.07</i>	<i>83.13</i>

it relative to DriveLM-CARLA in all context settings. This is likely due to the higher diversity in human answers obtained for DriveLM-nuScenes, as opposed to the rule-based generation in CARLA. On both datasets, DriveLM-Agent, which is fine-tuned on DriveLM, significantly outperforms BLIP-2 which is applied in a zero-shot manner. On both datasets and cases (zero-shot and fine-tuned), we see the potential of the graph-based structure. For the exact evaluation setting and a question-wise performance analysis, we refer to the supplementary.

5 Related Work

Generalization in Autonomous Driving. The inadequacy of generalization to the “long tail” of corner cases poses significant safety concerns to AD systems [10, 79, 80]. To tackle this issue, prior research primarily makes efforts in data-driven methods [1, 9, 29, 78, 88]. For example, TrafficSim [78] collects more data for safety-critical cases by simulation. An emerging direction involves leveraging semantic information to supervise the detection of unseen or anomalous objects [14, 24, 63, 89]. Even so, the zero-shot performance of AD systems is currently not satisfactory. In this paper, we bring a new approach towards better generalization: learning logical reasoning using Graph VQA.

Language-grounded Driving. Several concurrent methods attempt to incorporate multi-modal inputs into LLMs for AD tasks [11, 14, 28, 39, 57–59, 63, 72,

73, 81, 89, 91, 94, 102, 104]. Specifically, GPT-Driver [57] and LLM-Driver [11] encode the perceived scene state into prompts, relying on LLMs to formulate reasonable plans. DriveGPT4 [102] projects raw sensor data into tokens and utilizes LLMs for end-to-end prediction of control signals and explanations. Despite these preliminary attempts, there is untapped potential in addressing generalization in AD through LLMs. Our work combines VLMs with training over graph-structured QAs from DriveLM. This enables us to show benefits on zero-shot end-to-end planning, which was not demonstrated by these concurrent studies.

6 Discussion

Even though DriveLM exhibits promising generalization, there are concerning limitations of this work.

- **Efficiency Constraints.** Inheriting the drawbacks of LLMs, our baseline model suffers from long inference times, especially as we

require multiple rounds of predictions based on the graph structure (roughly $10\times$ slower than UniAD, as shown in Table 5). The core problem lies in the slow throughput of the current model (**8.5 tokens/s** in DriveLM-Agent), which may impact practical usage. However, LLMs have become orders of magnitude faster [74, 100] in the last months as this is a general topic of broad interest. We believe that the rapid progress in orthogonal research can alleviate this issue.

- **Driving-specific Inputs.** DriveLM-Agent directly applies the VLM’s vision module, taking a low-resolution front-view image as input. Driving-specific sensors such as LiDAR cannot be processed as well. This results in a lack of temporal information and 360-degree scene understanding. Extending DriveLM-Agent to images from multiple views is straightforward as the graph formulation allows various input frames for different nodes. We leave it for future work to explore options for multi-modal and multi-frame inputs.
- **Closed-loop Planning.** Our approach is currently evaluated under an open-loop scheme. In this setting, incorporating the ego vehicle’s status as input can significantly enhance the metrics, but its effectiveness may not translate well to the real world, and hence we only consider methods that do not do so. Extending our work to a closed-loop setting with an affordable budget in training time and computational cost is a promising direction to explore. With the usage of CARLA we provide a promising foundation for more research in the direction of closed-loop planning with VLMs.

Conclusion. We show how VLMs can be leveraged as end-to-end autonomous driving agents with improved generalization over task-specific driving stacks. For this we propose the task of Graph VQA together with new datasets and metrics. Equipped with these tools, we build a baseline approach that has a simple architecture and obtains promising results.

Table 5: Compared to UniAD-Single, DriveLM-Agent has fewer trainable parameters but a higher inference cost.

Method	#Params	#Trainable	FLOPs	FPS
UniAD-Single	131.9M	58.8M	1.7T	1.8
DriveLM-Agent	3.955B	12.9M	24.2T	0.16

Acknowledgements

The OpenDriveLab team is part of the Shanghai AI Lab and kindly supported by National Key R&D Program of China (2022ZD0160104) and NSFC (62206172). This paper is partially supported by the National Key R&D Program of China No.2022ZD0161000 and the General Research Fund of Hong Kong No.17200622 and 17209324. This work was also supported by the BMBF (Tübingen AI Center, FKZ: 01IS18039A), the DFG (SFB 1233, TP 17, project number: 276693517), and by EXC (number 2064/1 – project number 390727645). We thank the International Max Planck Research School for Intelligent Systems (IMPRS-IS) for supporting K. Renz and K. Chitta. Our gratitude goes to Tai Wang for the valuable feedback, Qingwen Bu for refining the figures, and Jiazhi Yang, Shen Yuan Gao, Yihang Qiu, Tianyu Li, Yunsong Zhou, Zetong Yang, Julian Zimmerlin for the fruitful discussion.

References

1. Akhauri, S., Zheng, L.Y., Lin, M.C.: Enhanced transfer learning for autonomous driving with systematic accident simulation. In: IROS (2020)
2. Anderson, P., Fernando, B., Johnson, M., Gould, S.: Spice: Semantic propositional image caption evaluation. In: ECCV (2016)
3. Atakishiyev, S., Salameh, M., Babiker, H., Goebel, R.: Explaining autonomous driving actions with visual question answering. arXiv preprint arXiv:2307.10408 (2023)
4. Beißwenger, J.: PDM-Lite: A rule-based planner for carla leaderboard 2.0. <https://github.com/OpenDriveLab/DriveLM/blob/DriveLM-CARLA/docs/report.pdf> (2024)
5. Besta, M., Blach, N., Kubicek, A., Gerstenberger, R., Gianinazzi, L., Gajda, J., Lehmann, T., Podstawski, M., Niewiadomski, H., et al.: Graph of Thoughts: Solving Elaborate Problems with Large Language Models. arXiv preprint arXiv:2308.09687 (2023)
6. Caesar, H., Bankiti, V., Lang, A.H., Vora, S., Liong, V.E., Xu, Q., Krishnan, A., Pan, Y., Baldan, G., Beijbom, O.: nuScenes: A multimodal dataset for autonomous driving. In: CVPR (2020)
7. Caesar, H., Kabzan, J., Tan, K.S., Fong, W.K., Wolff, E.M., Lang, A.H., Fletcher, L., Beijbom, O., Omari, S.: nuplan: A closed-loop ml-based planning benchmark for autonomous vehicles. In: CVPR Workshops (2021)
8. Chen, D., Koltun, V., Krähenbühl, P.: Learning to drive from a world on rails. In: ICCV (2021)
9. Chen, D., Krähenbühl, P.: Learning from all vehicles. In: CVPR (2022)
10. Chen, L., Wu, P., Chitta, K., Jaeger, B., Geiger, A., Li, H.: End-to-end autonomous driving: Challenges and frontiers. arXiv preprint arXiv:2306.16927 (2023)
11. Chen, L., Sinavski, O., Hünermann, J., Karnsund, A., Willmott, A.J., Birch, D., Maund, D., Shotton, J.: Driving with llms: Fusing object-level vector modality for explainable autonomous driving. arXiv preprint arXiv:2310.01957 (2023)
12. Chen, X., Jia, S., Xiang, Y.: A review: Knowledge reasoning over knowledge graph. Expert Syst. Appl (2020)

13. Chen, Y., Rohrbach, M., Yan, Z., Shuicheng, Y., Feng, J., Kalantidis, Y.: Graph-based global reasoning networks. In: CVPR (2019)
14. Chen, Y., Tonkens, S., Pavone, M.: Categorical traffic transformer: Interpretable and diverse behavior prediction with tokenized latent. arXiv preprint arXiv:2311.18307 (2023)
15. Chib, P.S., Singh, P.: Recent advancements in end-to-end autonomous driving using deep learning: A survey. IEEE T-IV (2023)
16. Chitta, K., Prakash, A., Geiger, A.: Neat: Neural attention fields for end-to-end autonomous driving. In: ICCV (2021)
17. Chitta, K., Prakash, A., Jaeger, B., Yu, Z., Renz, K., , Geiger, A.: Transfuser: Imitation with transformer-based sensor fusion for autonomous driving. IEEE T-PAMI (2023)
18. Database, A.L.: Incident 293: Cruise’s self-driving car involved in a multiple-injury collision at an san francisco intersection. <https://incidentdatabase.ai/cite/293/> (2022)
19. Dauner, D., Hallgarten, M., Geiger, A., Chitta, K.: Parting with misconceptions about learning-based vehicle motion planning. In: CoRL (2023)
20. Dewangan, V., Choudhary, T., Chandhok, S., Priyadarshan, S., Jain, A., Singh, A., Srivastava, S., Jatavallabhula, K., Krishna, M.: Talk2BEV: Language-enhanced bird’s-eye view maps for autonomous driving. arXiv preprint arXiv:2310.02251 (2023)
21. Ding, R., Zhang, C., Wang, L., Xu, Y., Ma, M., Zhang, W., Qin, S., Rajmohan, S., Lin, Q., Zhang, D.: Everything of thoughts: Defying the law of penrose triangle for thought generation. arXiv preprint arXiv:2311.04254 (2023)
22. Dosovitskiy, A., Ros, G., Codevilla, F., Lopez, A., Koltun, V.: CARLA: An open urban driving simulator. In: CoRL (2017)
23. Driess, D., Xia, F., Sajjadi, M.S.M., Lynch, C., Chowdhery, A., Ichter, B., Wahid, A., Tompson, J., Vuong, Q., et al.: PaLM-E: An embodied multimodal language model. In: ICML (2023)
24. Elhafsi, A., Sinha, R., Agia, C., Schmerling, E., Nesnas, I.A., Pavone, M.: Semantic anomaly detection with large language models. Auton. Robot (2023)
25. Floridi, L., Chiriatti, M.: GPT-3: Its nature, scope, limits, and consequences. MIND MACH (2020)
26. Gao, P., Han, J., Zhang, R., Lin, Z., Geng, S., Zhou, A., Zhang, W., Lu, P., He, C., Yue, X., Li, H., Qiao, Y.: LLaMA-Adapter v2: Parameter-efficient visual instruction model. arXiv preprint arXiv:2304.15010 (2023)
27. Groeger, J.A.: Understanding driving: Applying cognitive psychology to a complex everyday task. Routledge (2013)
28. Han, W., Guo, D., Xu, C.Z., Shen, J.: DME-Driver: Integrating human decision logic and 3d scene perception in autonomous driving. arXiv preprint arXiv:2401.03641 (2024)
29. Hanselmann, N., Renz, K., Chitta, K., Bhattacharyya, A., Geiger, A.: King: Generating safety-critical driving scenarios for robust imitation via kinematics gradients. In: ECCV (2022)
30. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: LoRA: Low-rank adaptation of large language models. In: CoRL (2021)
31. Hu, S., Chen, L., Wu, P., Li, H., Yan, J., Tao, D.: St-p3: End-to-end vision-based autonomous driving via spatial-temporal feature learning. In: ECCV (2022)
32. Hu, Y., Yang, J., Chen, L., Li, K., Sima, C., Zhu, X., Chai, S., Du, S., Lin, T., et al.: Planning-oriented autonomous driving. In: CVPR (2023)

33. Huang, W., Xia, F., Xiao, T., Chan, H., Liang, J., Florence, P., Zeng, A., Tompson, J., Mordatch, I., et al.: Inner monologue: Embodied reasoning through planning with language models. In: CoRL (2022)
34. Hudson, D.A., Manning, C.D.: Gqa: A new dataset for real-world visual reasoning and compositional question answering. In: CVPR (2019)
35. Ichter, B., Brohan, A., Chebotar, Y., Finn, C., Hausman, K., Herzog, A., Ho, D., Ibarz, J., Irpan, A., et al.: Do as i can, not as i say: Grounding language in robotic affordances. In: CoRL (2023)
36. Jaeger, B., Chitta, K., Geiger, A.: Hidden biases of end-to-end driving models. In: ICCV (2023)
37. Jaeger, B., Chitta, K., Geiger, A.: Hidden biases of end-to-end driving models. In: ICCV (2023)
38. Jelinek, F., Lafferty, J.D., Mercer, R.L.: Basic methods of probabilistic context free grammars. Springer, Berlin, Heidelberg (1992)
39. Jin, B., Liu, X., Zheng, Y., Li, P., Zhao, H., Zhang, T., Zheng, Y., Zhou, G., Liu, J.: Adapt: Action-aware driving caption transformer. In: ICRA (2023)
40. Karamcheti, S., Nair, S., Chen, A.S., Kollar, T., Finn, C., Sadigh, D., Liang, P.: Language-Driven representation learning for robotics. In: RSS (2023)
41. Kim, J., Misu, T., Chen, Y.T., Tawari, A., Canny, J.: Grounding human-to-vehicle advice for self-driving vehicles. In: CVPR (2019)
42. Kim, J., Rohrbach, A., Darrell, T., Canny, J., Akata, Z.: Textual explanations for self-driving vehicles. In: ECCV (2018)
43. Kojima, T., Gu, S.S., Reid, M., Matsuo, Y., Iwasawa, Y.: Large language models are zero-shot reasoners. In: NeurIPS (2022)
44. Lamb, A.M., ALIAS PARTH GOYAL, A.G., Zhang, Y., Zhang, S., Courville, A.C., Bengio, Y.: Professor forcing: A new algorithm for training recurrent networks. In: NeurIPS (2016)
45. Lavie, A., Agarwal, A.: METEOR: An automatic metric for MT evaluation with high levels of correlation with human judgments. In: ACL Workshop (2007)
46. Li, H., Sima, C., Dai, J., Wang, W., Lu, L., Wang, H., Zeng, J., Li, Z., Yang, J., Deng, H., et al.: Delving into the devils of bird’s-eye-view perception: A review, evaluation and recipe. IEEE T-PAMI (2023)
47. Li, J., Li, D., Savarese, S., Hoi, S.: BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In: ICML (2023)
48. Li, K., Chen, K., Wang, H., Hong, L., Ye, C., Han, J., Chen, Y., Zhang, W., Xu, C., Yeung, D.Y., et al.: CODA: A real-world road corner case dataset for object detection in autonomous driving. In: ECCV (2022)
49. Li, Z., Wang, W., Li, H., Xie, E., Sima, C., Lu, T., Qiao, Y., Dai, J.: Bevformer: Learning bird’s-eye-view representation from multi-camera images via spatiotemporal transformers. In: ECCV (2022)
50. Li, Z., Yu, Z., Lan, S., Li, J., Kautz, J., Lu, T., Alvarez, J.M.: Is ego status all you need for open-loop end-to-end autonomous driving? (2023)
51. Liang, J., Huang, W., Xia, F., Xu, P., Hausman, K., Ichter, B., Florence, P., Zeng, A.: Code as policies: Language model programs for embodied control. In: ICRA (2023)
52. Lin, C.Y.: Rouge: A package for automatic evaluation of summaries. In: ACL Workshop (2004)
53. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: ECCV (2014)
54. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. In: NeurIPS (2023)

55. Macadam, C.C.: Understanding and modeling the human driver. *Veh. Syst. Dyn* (2003)
56. Malla, S., Choi, C., Dwivedi, I., Choi, J.H., Li, J.: DRAMA: Joint risk localization and captioning in driving. In: *WACV* (2023)
57. Mao, J., Qian, Y., Zhao, H., Wang, Y.: GPT-Driver: Learning to drive with gpt. *arXiv preprint arXiv:2310.01415* (2023)
58. Mao, J., Ye, J., Qian, Y., Pavone, M., Wang, Y.: A language agent for autonomous driving. *arXiv preprint arXiv:2311.10813* (2023)
59. Marcu, A.M., Chen, L., Hünemann, J., Karnsund, A., Hanotte, B., Chidananda, P., Nair, S., Badrinarayanan, V., Kendall, A., et al.: LingoQA: Video question answering for autonomous driving. *arXiv preprint arXiv:2312.14115* (2023)
60. Marr, D.: *Vision: A computational investigation into the human representation and processing of visual information*. The MIT Press (2010)
61. OpenAI: OpenAI: Introducing ChatGPT. <https://openai.com/blog/chatgpt> (2022)
62. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., et al.: Training language models to follow instructions with human feedback. In: *NeurIPS* (2022)
63. Pan, C., Yaman, B., Nesti, T., Mallik, A., Allievi, A.G., Velipasalar, S., Ren, L.: VLP: Vision language planning for autonomous driving. *arXiv preprint arXiv:2401.05577* (2024)
64. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: *ACL* (2002)
65. Peng, Z., Wang, W., Dong, L., Hao, Y., Huang, S., Ma, S., Wei, F.: Kosmos-2: Grounding multimodal large language models to the world. *arXiv preprint arXiv:2306.14824* (2023)
66. Prakash, A., Chitta, K., Geiger, A.: Multi-modal fusion transformer for end-to-end autonomous driving. In: *CVPR* (2021)
67. Qian, T., Chen, J., Zhuo, L., Jiao, Y., Jiang, Y.G.: NuScenes-QA: A multi-modal visual question answering benchmark for autonomous driving scenario. *arXiv preprint arXiv:2305.14836* (2023)
68. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I.: Language models are unsupervised multitask learners. *OpenAI blog* (2019)
69. Rana, K., Haviland, J., Garg, S., Abou-Chakra, J., Reid, I., Suenderhauf, N.: SayPlan: Grounding large language models using 3d scene graphs for scalable robot task planning. In: *CoRL* (2023)
70. Renz, K., Chitta, K., Mercea, O.B., Koepke, A.S., Akata, Z., Geiger, A.: Plant: Explainable planning transformers via object-level representations. In: *CoRL* (2022)
71. Sachdeva, E., Agarwal, N., Chundi, S., Roelofs, S., Li, J., Dariush, B., Choi, C., Kochenderfer, M.: Rank2Tell: A multimodal driving dataset for joint importance ranking and reasoning. *arXiv preprint arXiv:2309.06597* (2023)
72. Seff, A., Cera, B., Chen, D., Ng, M., Zhou, A., Nayakanti, N., Refaat, K.S., Al-Rfou, R., Sapp, B.: MotionLM: Multi-agent motion forecasting as language modeling. In: *ICCV* (2023)
73. Shao, H., Hu, Y., Wang, L., Waslander, S.L., Liu, Y., Li, H.: LMDrive: Closed-loop end-to-end driving with large language models. *arXiv preprint arXiv:2312.07488* (2023)
74. Shi, D., Tao, C., Rao, A., Yang, Z., Yuan, C., Wang, J.: Crossget: Cross-guided ensemble of tokens for accelerating vision-language transformers (2023)
75. Shi, J., Zhang, H., Li, J.: Explainable and explicit visual reasoning over scene graphs. In: *CVPR* (2019)

76. Spelke, E.S., Kinzler, K.D.: Core knowledge. *Dev Sci* (2007)
77. Sun, P., Kretzschmar, H., Dotiwalla, X., Chouard, A., Patnaik, V., Tsui, P., Guo, J., Zhou, Y., Chai, Y., Caine, B., et al.: Scalability in perception for autonomous driving: Waymo open dataset. In: *CVPR* (2020)
78. Suo, S., Regalado, S., Casas, S., Urtasun, R.: Trafficsim: Learning to simulate realistic multi-agent behaviors. In: *CVPR* (2021)
79. Tampuu, A., Matiisen, T., Semikin, M., Fishman, D., Muhammad, N.: A survey of end-to-end driving: Architectures and training methods. *IEEE T-NNLS* (2020)
80. Teng, S., Hu, X., Deng, P., Li, B., Li, Y., Ai, Y., Yang, D., Li, L., Xuanyuan, Z., Zhu, F., et al.: Motion planning for autonomous driving: The state of the art and future perspectives. *IEEE T-IV* (2023)
81. Tian, X., Gu, J., Li, B., Liu, Y., Hu, C., Wang, Y., Zhan, K., Jia, P., Lang, X., Zhao, H.: DriveVLM: The convergence of autonomous driving and large vision-language models. *arXiv preprint arXiv:2402.12289* (2024)
82. Toomarian, N.B., Barhen, J.: Learning a trajectory using adjoint functions and teacher forcing. *Neural networks* (1992)
83. Touvron, H., Martin, L., Stone, K.R., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., et al.: Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288* (2023)
84. Treiber, M., Hennecke, A., Helbing, D.: Congested traffic states in empirical observations and microscopic simulations. *Physical Review E* **62**(2) (2000)
85. Treiber, M., Hennecke, A., Helbing, D.: Congested traffic states in empirical observations and microscopic simulations. *Physical review E* (2000)
86. Vedantam, R., Lawrence Zitnick, C., Parikh, D.: Cider: Consensus-based image description evaluation. In: *CVPR* (2015)
87. Wang, H., Li, T., Li, Y., Chen, L., Sima, C., Liu, Z., Wang, B., Jia, P., Wang, Y., Jiang, S., et al.: OpenLane-V2: A topology reasoning benchmark for unified 3d HD mapping. In: *NeurIPS Datasets and Benchmarks* (2023)
88. Wang, J., Pun, A., Tu, J., Manivasagam, S., Sadat, A., Casas, S., Ren, M., Urtasun, R.: Advsim: Generating safety-critical scenarios for self-driving vehicles. In: *CVPR* (2021)
89. Wang, T.H., Maalouf, A., Xiao, W., Ban, Y., Amini, A., Rosman, G., Karaman, S., Rus, D.: Drive anywhere: Generalizable end-to-end autonomous driving with multi-modal foundation models. *arXiv preprint arXiv:2310.17642* (2023)
90. Wang, W., Chen, Z., Chen, X., Wu, J., Zhu, X., Zeng, G., Luo, P., Lu, T., Zhou, J., Qiao, Y., Dai, J.: VisionLLM: Large language model is also an open-ended decoder for vision-centric tasks. *arXiv preprint arXiv:2305.11175* (2023)
91. Wang, W., Xie, J., Hu, C., Zou, H., Fan, J., Tong, W., Wen, Y., Wu, S., Deng, H., et al.: DriveMLM: Aligning multi-modal large language models with behavioral planning states for autonomous driving. *arXiv preprint arXiv:2312.09245* (2023)
92. Wang, X., Wang, D., Xu, C., He, X., Cao, Y., Chua, T.S.: Explainable reasoning over knowledge graphs for recommendation. In: *AAAI* (2019)
93. Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., Narang, S., Chowdhery, A., Zhou, D.: Self-Consistency improves chain of thought reasoning in language models. In: *ICLR* (2023)
94. Wayve: Lingo-1. <https://wayve.ai/thinking/lingo-natural-language-autonomous-driving/> (2023)
95. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., Zhou, D.: Chain-of-thought prompting elicits reasoning in large language models. In: *NeurIPS* (2022)

96. Wen, C., Lin, J., Qian, J., Gao, Y., Jayaraman, D.: Keyframe-focused visual imitation learning. In: ICML (2021)
97. Wu, D., Han, W., Wang, T., Dong, X., Zhang, X., Shen, J.: Referring Multi-Object tracking. In: CVPR (2023)
98. Wu, D., Han, W., Wang, T., Liu, Y., Zhang, X., Shen, J.: Language prompt for autonomous driving. arXiv preprint arXiv:2309.04379 (2023)
99. Wu, P., Jia, X., Chen, L., Yan, J., Li, H., Qiao, Y.: Trajectory-guided control prediction for end-to-end autonomous driving: A simple yet strong baseline. In: NeurIPS (2022)
100. Xiao, G., Lin, J., Seznec, M., Wu, H., Demouth, J., Han, S.: SmoothQuant: Accurate and efficient post-training quantization for large language models. In: Proceedings of the 40th International Conference on Machine Learning (2023)
101. Xinpeng, D., Jinahua, H., Hang, X., Xiaodan, L., Xu, H., Wei, Z., Xiaomeng, L.: Holistic autonomous driving understanding by bird’s-eye-view injected multi-modal large models. In: CVPR (2024)
102. Xu, Z., Zhang, Y., Xie, E., Zhao, Z., Guo, Y., Wong, K.K., Li, Z., Zhao, H.: DriveGPT4: Interpretable end-to-end autonomous driving via large language model. arXiv preprint arXiv:2310.01412 (2023)
103. Yao, S., Yu, D., Zhao, J., Shafran, I., Griffiths, T.L., Cao, Y., Narasimhan, K.: Tree of Thoughts: Deliberate problem solving with large language models. arXiv preprint arXiv:2305.10601 (2023)
104. Yuan, J., Sun, S., Omeiza, D., Zhao, B., Newman, P., Kunze, L., Gadd, M.: RAG-Driver: Generalisable driving explanations with retrieval-augmented in-context learning in multi-modal large language model. arXiv preprint arXiv:2402.10828 (2024)
105. Zareian, A., Rosa, K.D., Hu, D.H., Chang, S.F.: Open-vocabulary object detection using captions. In: CVPR (2021)
106. Zhang, R., Han, J., Zhou, A., Hu, X., Yan, S., Lu, P., Li, H., Gao, P., Qiao, Y.: LLaMA-Adapter: Efficient fine-tuning of language models with zero-init attention. arXiv preprint arXiv:2303.16199 (2023)
107. Zhou, X., Liu, M., Zagar, B.L., Yurtsever, E., Knoll, A.C.: Vision language models in autonomous driving and intelligent transportation systems. arXiv preprint arXiv:2310.14414 (2023)
108. Zitkovich, B., Yu, T., Xu, S., Xu, P., Xiao, T., Xia, F., Wu, J., Wohllhart, P., Welker, S., et al.: RT-2: Vision-language-action models transfer web knowledge to robotic control. In: CoRL (2023)

Overview

In the appendices below, we first delve deeper into various discussions, along with additional details around the annotation process of DriveLM-nuScenes & -CARLA, GVQA metrics, context setting & trajectory tokenization in DriveLM-Agent, and more ablation results of DriveLM-Agent on DriveLM-nuScenes and Waymo. Finally, we provide additional results and visualizations that further complement the findings from the main text.

For readers who want to focus on specific topics, we provide a summary below:

Appendix A – *Motivating Questions*

We index a list of “motivating” questions that may arise from reading the main text and that we expand on further here (*e.g.*, “why adapt general VLMs to driving”). These questions are open to be explored and thus our answers here are *intuitive* and *empirical*.

Appendix B – *DriveLM-nuScenes*

We provide the DriveLM-nuScenes dataset composition, introduce the detailed annotation pipeline and conduct statistics of QA categories.

Appendix C – *DriveLM-CARLA*

We provide a detailed description of the PDM-Lite expert, composition of the dataset, how the graph looks, and explain the data generation and annotation process.

Appendix D – *DriveLM-Metrics*

We explain the details of the metrics for each task in the GVQA, illustrate their differences, and provide the reasons for proposing GPT-score as the main metric used in the P_{1-3} VQA tasks.

Appendix E – *DriveLM-Agent*

We introduce the detailed design of the prompting with context and the trajectory tokenization, including the differences in context during training and evaluation, the pattern for the trajectory as a sentence, and the hyperparameters in the tokenizer.

Appendix F – *Experiments*

We provide more experiments, including the generalization to unseen objects, effects of more context design on the zero-shot ability, the performance with more conventional VQA metrics and the model efficiency comparison.

Appendix G – *Qualitative Results*

We show qualitative examples of the context, questions, and answers on nuScenes, Waymo, and CARLA. Additionally, we contrast predicted and ground truth answers together with their SPICE and GPT Score on nuScenes to provide some intuition for those metrics.

Appendix H – *Other Related Work*

We provide more related works from two new perspectives. One is reasoning over graph structure which is similar to our idea of graph-structure reasoning, the other is more vision-language benchmarks for autonomous driving.

A Motivating Questions

Q1. *In what situations could we expect planning with VLMs to outperform conventional end-to-end autonomous driving?*

One of the key challenges of autonomous driving is to generalize to the long-tail of scenarios, that are rarely encountered but have critical importance. Considering the large-scale pre-training of VLMs, their acquired knowledge of the world, and the reasoning ability of the LLM, it is anticipated that planning with VLMs will work better in situations that are novel or unseen in the context of driving scenarios but encountered during pre-training in unrelated contexts.

Q2. *Why adapt general VLMs to driving rather than adding language inputs to driving-specific models?*

General VLMs benefit from billion-scale pre-training data for vision-language tasks extracted from the internet, which can be adapted to the driving domain through fine-tuning on small autonomous driving datasets like DriveLM. Conversely, driving-specific models are only pre-trained on small autonomous driving datasets, and adding language inputs to these with data from outside the self-driving domain is non-trivial. Combining the advantages of VLMs and driving-specific models is however an interesting direction to explore.

Q3. *Can open-loop evaluation of planning provide meaningful results?*

When performing open-loop evaluation, providing the ego history as an input to the planning module prevents fair comparisons, as this signal alone is sufficient for achieving low errors on existing benchmarks. DriveLM alleviates this issue by evaluating key frames, where the intention of the ego vehicle changes, and the ego history is not strongly indicative of the future behavior or motion. Additionally, we consider baselines in our analysis that do not input the ego history to the planning module. Finally, we introduce DriveLM-CARLA as a means to show closed-loop planning results in the future.

Q4. *Why are there currently no closed-loop planning results on CARLA?*

Running 4B parameter models at 20 FPS as required by CARLA needs more engineering effort. This could be solved by using distillation, quantization, and caching techniques in LLM inference. Another approach would be to execute only the final motion stage of DriveLM-Agent at 20 FPS, while the other GVQA stages are executed at a lower frame rate.

Q5. *Is DriveLM-Agent efficient enough to be applicable to real-world autonomous driving?*

We comment on the runtime of DriveLM-Agent in Table 5 in the main paper. Without any optimization, the approach is around 1 order of magnitude slower than UniAD. However, with the optimizations proposed for closed-loop results on CARLA (see Q4), practical applications of VLMs in driving should be possible.

Q6. *What is the trade-off between long inference time and generalization?*

LLMs have become orders of magnitude faster in the last years as they are a general topic of broad interest, e.g. [100]. Similarly, recent works show BLIP-2 can be run 40% faster while maintaining performance [74]. Furthermore, most AD research begins with systems that are not real-time, which are later optimized by practitioners. We admit that runtime efficiency is a limitation of DriveLM but we hope that the rapid progress in orthogonal research can alleviate this issue, which is out of the scope of our project.

Q7. *Why is VQA more suited than alternative techniques to train internet-scale models (such as generative modeling) for the downstream application of autonomous driving?*

Both perception and planning in driving require reasoning and involve zero-shot generalization. VLMs potentially have the reasoning ability inherited from LLMs, making VQA a promising direction for bringing the benefits of web-scale training to autonomous driving.

Q8. *Do today’s VLMs understand and reason about the visual world as well as LLMs understand text-based worlds?*

This is not known but deserves to be explored. VLMs approach the problem of generalization in a data-driven way, which has been proved successful repeatedly on other tasks.

Q9. *Why does the proposed graph reasoning scheme not provide very strong improvements in VQA?*

It is possible that the simple prompting scheme, relatively small base VLMs, or insufficiently strong logical dependencies in the dataset (or a combination of these factors) contribute to the lack of major improvements. DriveLM-CARLA provides a platform to carefully study these factors and inform the annotation of future datasets for GVQA.

Q10. *What questions should be asked when collecting DriveLM?*

Determining the right questions is a critical aspect of the DriveLM system. We are unable to recommend a detailed protocol for this in our work, which is a pioneering study for driving with VLMs. This relies on domain expertise, which we aim to address in future work. See Table 3 in the main paper for a question-wise analysis of our chosen protocol.

Q11. *Is generalization in automobiles an oversold topic?*

Driving scenarios predominantly consist of vehicles, pedestrians, and cyclists. However, real-world AD is a safety-critical application that must handle the long tail of rare objects to avoid accidents in commercial deployment [18]. Furthermore, generalization to novel objects is an active and growing research field as shown in [48, 105]. While it is true that most of the objects in autonomous driving are limited, the tiny percentage of rare objects are in fact the main existing

barrier to the commercialization of AD, making generalization an important feature.

Q12. *Why not use video inputs?*

It has been observed that in AD multi-frame inputs do not always lead to improved performance due to causal confusion [96], motivating the use of single frames in several prior works [37, 99]. However, we agree that in the long term, multi-frame models are desirable. As a proof of concept, we provide the results for one (adapted from LLaMA-Adapter-V2 [26]) in Table 13. Multi-frame inputs give a slight improvement.

Q13. *What is the technical novelty in this paper?*

The young field of driving with language still lacks a standardized dataset, task and evaluation framework. Our paper fills this gap. We believe that these are as valuable as proposing a new method. Therefore, we do not make any claims regarding algorithmic novelty, intentionally constructing simple baselines. Our technical novelty lies in the formulation of the task and preparation of data suitable for adapting general VLMs to the driving application, which we show to have promising results even in combination with existing models.

B DriveLM-nuScenes

In this section, we introduce the details of DriveLM-nuScenes, including the dataset composition, collection methodology, and statistics.

B.1 Dataset Composition

DriveLM-nuScenes comprises **a training set of 4072 frames and a validation set of 799 frames**, consisting of scene-level descriptions and frame-level QA accompanied by 2D bounding boxes within multi-view images from the nuScenes dataset. The scene-level description delineates the behavior of the ego vehicle throughout the entire video clip. The frame-level QA encompasses three distinct categories: perception, prediction, and planning.

- **Perception** involves queries related to the thorough examination of the entire frame. Apart from several questions in this question set that are manually annotated, we design prompts to generate questions about the observational facets of objects within the scene, leveraging ground truth from nuScenes [6] and OpenLane-V2 [87].
- **Prediction** encompasses a series of inquiries regarding the projection of the forthcoming state of key objects and the ego vehicle in the current frame, and the underlying reasoning process behind the prediction. Because the predictions are intricate and challenging, we *manually annotate* the answers.

- **Planning** contains questions related to planning subsequent actions of the ego vehicle in the current frame. As “Planning” is the same challenging as “prediction”, we design the prompt for the reasoning process and *manually annotate* the answers to the questions.

For the key objects referred to in the QA, we encode them as **c tags** in the format $\langle c, CAM, x, y \rangle$, where c is the identifier, CAM indicates the camera where the object’s center point is situated, and x, y represent the horizontal and vertical coordinates of the 2D bounding box in the respective camera’s coordinate system. We also provide a dictionary in each key frame, recording more basic information about the key objects such as the size of the bounding box, the category, the moving state, and the visual description. The overview of data organization forms is shown in Fig. 5.

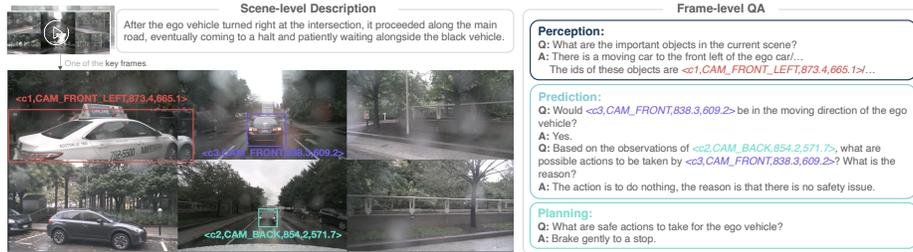


Fig. 5: Overall Composition of DriveLM-nuScenes. The dataset comprises scene-level descriptions and frame-level QA, which can be divided into three parts: *Perception*, *Prediction*, and *Planning*. Objects are encoded using *c tags*, which contain identifiers, camera affiliations, and center coordinates of its 2D bounding box in the corresponding camera frame.

B.2 Collection Methodology

During the annotation process, we employ individuals with driving experience for the labeling task. We provide annotators with the stitched results from the six cameras of nuScenes as source data. As shown in Fig. 6 (left), we divide the annotation process into three steps: selecting key frames from video clips, choosing key objects within these key frames, and subsequently annotating the frame-level QAs in the key frames. Following this, we conduct multiple rounds of quality checks to ensure the data reliability and perform post-processing procedures on the qualified data as shown in Fig. 6 (right). The specific details of this pipeline will be introduced below.

Key Frame Selection. In this process, we ask annotators to review the entire video clip to pinpoint key frames rich in scene information and potentially indicative of future state changes. Simultaneously, annotators are instructed to

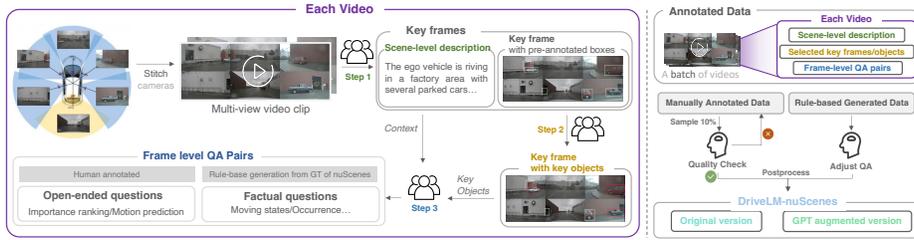


Fig. 6: (Left) Pipeline of the three-steps annotation process. For each video, we ask the annotators to annotate the key frames, key objects, and QA attributes step by step. **(Right) The quality check and post-processing progress.** We divided the annotated data into batches, where each batch contains 8 video clips and their related annotations. We conduct rigorous quality checks, and after the post-processing, we finally get two versions of our DriveLM-nuScenes dataset.

label the ego vehicle’s behavior throughout the video clip. This segment serves as the foundation for our scene-level description.

Key Object Selection. In this annotation step, we instruct annotators to identify objects in key frames that are relevant to the ego vehicle’s driving, denoted as key objects. To ensure accuracy, we provide pre-annotated bounding boxes based on ground truth categories from nuScenes [6]. Annotators also have the flexibility to designate objects not present in the ground truth as key objects if they are deemed significant.

QA Labeling. In the QA labeling process, we have two sets of questions, factual questions and open-ended questions. For the factual questions, we generate the answers with a rule-based method. For the open-ended questions, we instruct annotators to manually annotate the meticulously designed questions. Options are provided for most manually annotated questions, and we include an "Other - Fill in the Blank" option for answer choices in such cases to ensure flexibility. We have also incorporated free-form questions, allowing annotators to generate their own inquiries about the current frame.

Quality Check. We prioritize the quality of our data. In addition to establishing clear criteria and implementing autonomic checking strategies at each annotation step, we conduct rigorous manual quality checks. We organize the final data into batches, with each batch comprising 8 video clips, along with their scene-level descriptions, key frames with key objects selected from the 8 video clips, and corresponding QA pairs for each key frame. We provide explicit standards to quality check inspectors, instructing them to assess data eligibility based on these criteria. For manually annotated data, if the accuracy of the manual annotations falls below expectations for a particular batch, we compile feedback on the issues encountered and request annotators to re-annotate the entire batch. For data generated from ground truth, we instruct quality inspectors to manually adjust the inconsistent or unreasonable QA pairs.

Post Processing. Since our annotators are Chinese speakers, we need to translate the labeled data into English after obtaining it. Initially, we establish map-

pings between Chinese and English using a vocabulary. For texts that are not successfully mapped, we utilize GPT-3.5 for translation, and then perform manual checks and corrections on the GPT outputs. We also provide a version augmented by GPT-3.5, utilizing the prompt as shown in Table 6.

```
Messages = [
  { "role": "system", "content": f""" You are an English improver.
    """ },
  { "role": "user", "content": f""" I have a question and answer
    that I need you to help me modify and embellish, please make a few
    simple changes to the content in written language and keep the meaning
    same, you only need to answer the changes to: {QA}"""}]
```

Table 6: Prompt for GPT-refined version of DriveLM-nuScenes. We try 50 different prompt and select this pattern as the final one to do the refinement.

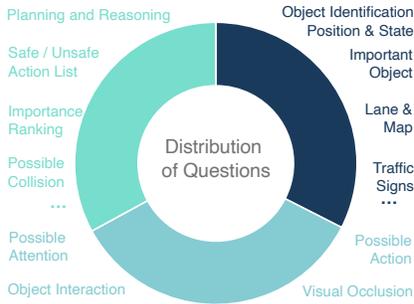


Fig. 7: Question Distribution of Perception, Prediction and Planning: The questions in our dataset cover various specific aspects of driving tasks, generally categorized into perception, prediction, and planning. Most of these questions are annotated by human annotators, making this a suitable proxy for human-like driving reasoning.

B.3 Statistics and Facts

In this section, we conduct a distribution analysis of our DriveLM-nuScenes QA categories at both the task level and object level. Additionally, for the task level, we provide the templates for all our QA under this classification criterion. The results indicate the richness of our QA categories, covering various aspects of autonomous driving. Moreover, the abundance of logical relationships is sufficient to construct a graph-structured QA.

Task level. Our DriveLM-nuScenes orchestrates a benchmark that encompasses various aspects of autonomous driving, connecting the whole stages of the human driving logic. To delve deeper into this aspect, we present the detailed QA types

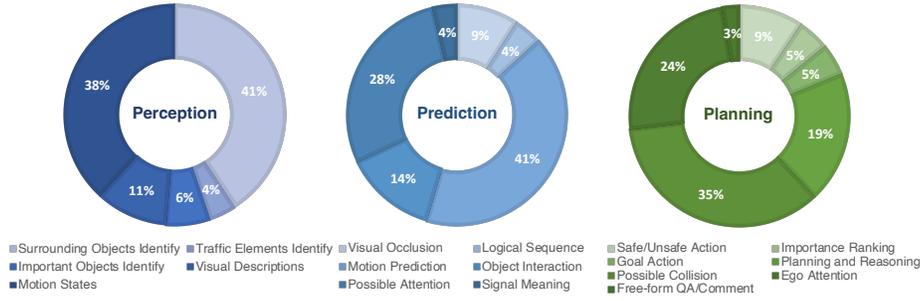


Fig. 8: The distribution of question types according to different tasks in DriveLM-nuScenes. We categorize questions into perception, prediction, and planning tasks, each further subdivided into more specific question types.

distribution at task level in Fig. 8. For a better understanding, we also provide examples of QA templates in all of the P3 stages in Table 7.

Perception

Surrounding Objects Identify

Q: Please describe the current scene.
 A: There are two moving cars behind the ego car and two barriers in front of it.

Q: What are objects to the front left/back right/... of the ego car?
 A: There are two barriers to the front left of the ego car.

Q: Are there traffic cones/moving cars/... to the front right/back left/... of the ego car?
 A: No.

Traffic Elements Identify

Q: Is there any traffic element in the front view?
 A: Yes, there are some traffic elements in the front view.

Q: Identify all the traffic elements in the front view, categorize them, determine their status, and predict the bounding box around each one. The output should be a list formatted as (c, s, x1, y1, x2, y2), where c represents the category, s denotes the status, and x1, y1, x2, y2 are the offsets of the top-left and bottom-right corners of the box relative to the center point.
 A: There are three traffic elements in the front view. The information of these traffic elements are [(road sign, go straight, 907.58, 590.67, 992.54, 630.95)...].

Important Objects Identify

Q: What are the important objects in the current scene? Those objects will be considered for the future reasoning and driving decision.

A: There is a parked truck to the back of the ego car... The ids of these objects are <c1,CAM_BACK,827.5,484.2>...

Q: What is the relative positioning of the important objects in the current scene?

A: <c3,CAM_FRONT,689.2,527.5> is to the front of <c1,CAM_BACK,827.5,484.2>...

Q: Which lanes are each important object on in the scene?

A: <c2,CAM_FRONT,820.8,473.3> is on the ego lane...

Visual Description

Q: What is the visual description of <c2,CAM_FRONT_LEFT,415.8,580.8>/...?

A: Pedestrian riding a bicycle.

Motion State

Q: What is the status of the cars/pedestrians/... that are to the front/front right/... of the ego car?

A: Many cars are parked.

Q: What is the observed status of object <c1,CAM_FRONT,920.0,509.2>/...?

A: Moving.

Q: What is the moving status of object <c1,CAM_FRONT,920.0,509.2>/...?

A: Going ahead.

Prediction

Visual Occlusion

Q: Which object is most likely to be occluded by <c1,CAM_FRONT,707.5,472.5>/...? Would this object affect the ego vehicle? Based on this object, what action of the ego vehicle is dangerous?

A: The object in front of <c1,CAM_FRONT,840.8,507.5>, yes, accelerating forward.

Logical Sequence

Q: What object should the ego vehicle notice first when the ego vehicle is getting to the next possible location? What is the state of the object that is first noticed by the ego vehicle and what action should the ego vehicle take? What object should the ego vehicle notice second when the ego vehicle is getting to the next possible location? What is the state of the object perceived by the ego vehicle as second and what action should the ego vehicle take? What object should the ego vehicle notice third? What is the state of the object perceived by the ego vehicle as third and what action should the ego vehicle take?

A: Firstly notice that <c2,CAM_FRONT,514.7,462.2>, the state of it is traffic sign, so the ego vehicle should slow down and go ahead. Secondly

notice that $\langle c3, \text{CAM_FRONT}, 950.3, 613.1 \rangle$, the state of it is traffic sign, so the ego vehicle should slow down and go ahead. Thirdly notice that $\langle c1, \text{CAM_FRONT}, 707.5, 472.5 \rangle$, the state of it is going ahead, so the ego vehicle should slow down and go ahead.

Motion Prediction

Q: Would $\langle c1, \text{CAM_FRONT}, 920.0, 509.2 \rangle / \dots$ be in the moving direction of the ego vehicle?

A: Yes.

Q: What is the future state of $\langle c1, \text{CAM_FRONT}, 920.0, 509.2 \rangle / \dots$?

A: Keep going straight.

Q: Will $\langle c2, \text{CAM_FRONT}, 1223.3, 598.3 \rangle / \dots$ be in the moving direction of $\langle c1, \text{CAM_BACK}, 514.2, 503.3 \rangle / \dots$?

A: No.

Object Interaction

Q: Will $\langle c2, \text{CAM_FRONT}, 1223.3, 598.3 \rangle / \dots$ change its motion state based on $\langle c1, \text{CAM_BACK}, 514.2, 503.3 \rangle / \dots$?

A: No.

Q: Based on the observations of $\langle c1, \text{CAM_BACK}, 514.2, 503.3 \rangle / \dots$, what are possible actions to be taken by $\langle c2, \text{CAM_FRONT}, 1223.3, 598.3 \rangle / \dots$? What is the reason?

A: The action is to keep going at the same speed, the reason is there is no safety issue.

Q: Based on the observation of $\langle c4, \text{CAM_FRONT}, 1071.2, 346.2 \rangle / \dots$, what actions may $\langle c1, \text{CAM_FRONT}, 1126.7, 515.0 \rangle / \dots$ take?

A: The action is to keep going at the same speed, the reason is there is no safety issue.

Possible Attention

Q: In this scenario, what object is most likely to consider $\langle c3, \text{CAM_FRONT}, 400.1, 717.2 \rangle / \dots$?

A: The ego vehicle.

Q: Would $\langle c1, \text{CAM_BACK}, 514.2, 503.3 \rangle / \dots$ take $\langle c3, \text{CAM_FRONT}, 400.1, 717.2 \rangle / \dots$ into account?

A: No.

Q: What object would consider $\langle c1, \text{CAM_FRONT}, 985.8, 516.7 \rangle / \dots$ to be most relevant to its decision?

A: The ego vehicle.

Q: Except for the ego vehicle, what object would consider $\langle c1, \text{CAM_FRONT}, 985.8, 516.7 \rangle / \dots$ to be most relevant to its decision?

A: $\langle c2, \text{CAM_FRONT}, 1217.5, 511.7 \rangle$.

Signal Meaning

Q: What does $\langle c2, \text{CAM_BACK_LEFT}, 400.8, 654.2 \rangle / \dots$ mean?

A: No entry.

Q: What kind of traffic sign is <c2,CAM_BACK_LEFT,400.8,654.2>/...?

A: Traffic cone.

Planning

Safe/Unsafe Action

Q: In this scenario, what are safe actions to take for the ego vehicle?

A: Decelerate gradually without braking, keep going at the same speed.

Q: In this scenario, what are dangerous actions to take for the ego vehicle?

A: Accelerate and go ahead, brake suddenly, drive backward, turn right.

Importance Ranking

Q: What is the priority of the objects that the ego vehicle should consider (in descending order)

A: <c2,CAM_FRONT,514.7,462.2>, <c3,CAM_FRONT,950.3,613.1>, <c1,CAM_FRONT,707.5,472.5>.

Goal Action

Q: What is the target action of the ego vehicle?

A: Go straight.

Planning and Reasoning

Q: What actions could the ego vehicle take based on <c1,CAM_FRONT,920.0,509.2>/...? Why take this action and what's the probability?

A: The action is to decelerate gradually without braking, the reason is to keep a safe distance, high.

Q: Based on <c3,CAM_FRONT,1591.1,441.8>/... in this scene, what is the most possible action of the ego vehicle?

A: Decelerate gradually without braking.

Possible Collision

Q: What is the probability of colliding with <c1,CAM_FRONT,920.0,509.2>/... after the ego vehicle goes straight and keeps the same speed/accelerates and goes straight/...?

A: Low.

Q: What actions taken by the ego vehicle can lead to a collision with <c1,CAM_FRONT,920.0,509.2>/...?

A: Accelerate and go straight.

Ego Attention

Q: What is the traffic signal that the ego vehicle should pay attention to?

A: None.

Q: Is <c1,CAM_FRONT,920.0,509.2>/... an object that the ego vehicle should consider in the current scene?

A: Yes.

<p>Q: Is it necessary for the ego vehicle to take <c3,CAM_FRONT,400.1,717.2>/... into account?</p> <p>A: Yes.</p> <p>Free-form QA/Comment</p> <p>Q: What impact does this situation have on driving vehicles?</p> <p>A: The road scene is complex, please slow down.</p> <p>Q: What's your comment on this scene?</p> <p>A: Pedestrians at the intersection, please be careful and give way.</p> <p>...</p>

Table 7: Question templates of DriveLM-nuScenes at task level. The categories in the table correspond to those in Fig. 8.

Object level. We also conduct some statistics at the object level since QAs in our DriveLM-nuScenes revolve around key objects. Fig. 9 (left) shows the distribution of our key object types. Given the substantial differences in questions associated with traffic elements compared to other categories, we separately conduct statistics for QA types related to traffic elements and the remaining categories. The results are depicted in Fig. 9 (right).

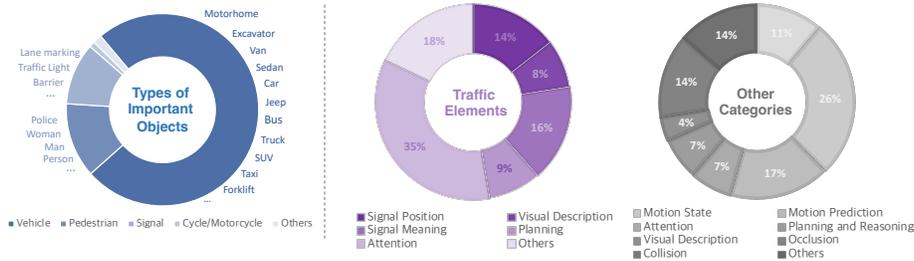


Fig. 9: (Left) The distribution of key objects in DriveLM-nuScenes. The sub-categories are extracted from the visual description. **(Right) The distribution of question types related to different key objects in DriveLM-nuScenes.** Since the questions associated with traffic elements differ significantly from other categories, we separately conduct statistics for QA types related to traffic elements and the remaining categories.

C DriveLM-CARLA

In this section, we introduce the details of DriveLM-CARLA, including the dataset composition and collection methodology.

C.1 Dataset Composition

DriveLM-CARLA consists of automatically generated frame-level question-answer pairs that are structured with an interconnected graph. The graph structure can be seen in Fig. 10. In the current version, the dataset consists of questions about the road layout, stop signs, traffic lights, and vehicles. In future versions, the dataset can be extended to more categories like static objects, weather, traffic signs, and others.

Utilizing the driving simulator CARLA for the data generation process allows for scalable annotations and data without any manual effort involved. Additionally, the dataset supports a variety of sensor outputs from CARLA, including semantic segmentation, depth map, LiDAR, and others, which can be employed to train different network architectures. Each question within the graph is designed in a way to facilitate situational reasoning, which could be instrumental in answering subsequent questions. As with DriveLM-nuScenes, each question can be categorized into perception, prediction, or planning. For each QA-pair, besides the corresponding question and answer, we also save the object ID in case the QA-pair is about an object. This ID is consistent over time, enabling object tracking and temporal reasoning in future studies. In addition, relationships to parent and child questions within the graph are documented to allow efficient traversal of the graph.

C.2 Expert Algorithm: PDM-Lite

While previous CARLA expert algorithms like the privileged rule-based expert used by TransFuser++ [36] can only solve the scenarios implemented in CARLA Leaderboard 1.0, PDM-Lite [4] is designed to tackle all 38 scenarios present in Leaderboard 2.0. It consists of six distinct stages, summarized in Fig. 11.

Path Planning. First, PDM-Lite creates a dense path in the form of spatially equidistant points using an A* planning algorithm, given sparse target points (up to 200m apart) by the leaderboard module. The plan is based on the HD map of the town, provided by CARLA. Furthermore, we add information such as speed limits and distances to the next traffic light/stop sign to different sections of this route that is to be traversed. To handle scenarios that require leaving the default path (e.g. *Accident*, *ConstructionObstacle*, *ParkedObstacle*, *VehicleOpensDoor*, *AccidentTwoWays*, *ConstructionObstacleTwoWays*, *ParkedObstacleTwoWays*, *VehicleOpensDoorTwoWays*, *YieldToEmergencyVehicle*), a short segment of the route where the scenario will be spawned is shifted laterally towards an adjacent lane.

Agent Forecasting. PDM-Lite generates a forecast of dynamic agents for a horizon of 2s with a temporal resolution of 20Hz. Since we do not know other actors' paths and controls in advance, we assume they maintain their previous controls and apply similar ones in the near future. Using a kinematic bicycle model with parameters taken from [8], PDM-Lite forecasts other agents' motion using this constant action assumption, similar to the expert of [36]. We only consider actors closer than 50m to the ego agent for forecasting.

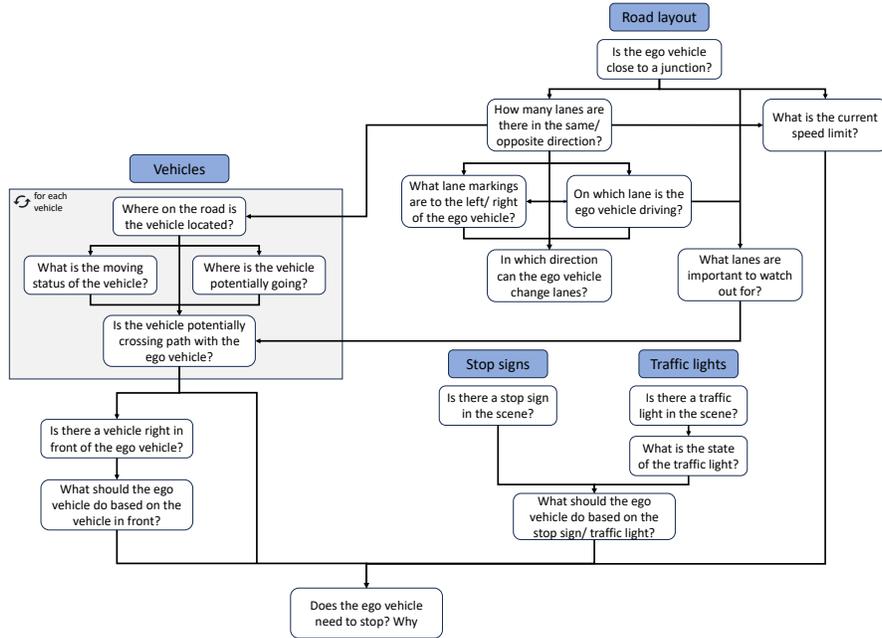


Fig. 10: Detailed flow of CARLA graph. We show the full graph of DriveLM-CARLA. The graph consists of questions and answers about the road layout, traffic lights, stop signs, and vehicles.

IDM Target Speed. We generate a target speed proposal using the Intelligent Driver Model [85]. Specifically, while the original IDM selects only one vehicle as a leading actor, we instead iteratively apply it to all vehicles, pedestrians, non-green traffic lights, and stop signs intersecting the path ahead of the ego vehicle. Once the leading actors are selected, we determine a target speed for each actor using the parameters summarized in Table 8. We use the minimum speed value obtained as the final target speed proposal for that timestep.

Simulation. We simulate the trajectory with the proposed target speed for 2s at 20Hz by alternatively applying the longitudinal controller (described in the following) and a kinematic bicycle model. Thereby, the proposal is converted into an actual expected sequence of ego-vehicle bounding boxes in closed-loop.

Scoring. Having forecasted the bounding boxes for all actors, we can now check for bounding box intersections between the simulated ego vehicle and other vehicles. We score the ego vehicle’s motion accordingly: if we detect an intersection with a non-leading and non-rear-end vehicle, we reject the IDM target speed proposal, and instead set the target speed to zero.

Controllers. Controlling the vehicle requires three values: steer, throttle, and brake. The steering value can be directly estimated from the ego vehicle’s location, velocity, and path. We use a lateral PID controller for doing so, similar

Table 8: Target Speed Proposal by IDM. PDM-Lite uses IDM to select a target speed. The parameters for the desired net distance and desired time headway differ with respect to the type of the leading actor: vehicles, bicycles / stop signs / traffic lights / walkers / static objects (e.g. construction signs).

Parameter	Value	Description
v_0	$0.72 v_{\text{lane}}$	Desired velocity. 72% of the speed limit
s_0	4.0/2.0/6.0/4.0/2.0 m	Desired net distance to the leading agent
T	0.25/0.1/0.1/0.25/0.1 s	Desired time headway to leading agent
a	24.0 ms^{-2}	Maximum acceleration of ego
$b_{v \leq 6.02}$	8.7 ms^{-2}	Maximum deceleration of ego if $v \leq 6.02$
$b_{v > 6.02}$	3.72 ms^{-2}	Maximum deceleration of ego if $v > 6.02$
δ	4.0	Acceleration exponent

to [36], which minimizes the angle to a selected point along the path ahead. For the throttle and brake predictions, we employ a linear regression model using features extracted based on the current speed and target speed.

Results. The performance of PDM-Lite is evaluated on the 20 official validation routes of Leaderboard 2.0. The results of three different seed evaluations are presented in Table 9. Routes 0-9 and 10-19 are identical except for weather conditions, providing an additional measure of performance variance. Since PDM-Lite utilizes privileged information, its variance in performance is not due to the differing weather parameters, but rather the random initialization of the surrounding traffic.

C.3 Collection Methodology

In this section, we provide details about the data collection and the annotation process.

Simulator settings. We utilize the CARLA Simulator (version 0.9.14) with Leaderboard 2.0 [22] to generate our dataset. Leaderboard 2.0 introduces two new large maps along with a suite of new scenarios, enhancing the diversity of the training and evaluation environments. Town 12 serves as the training

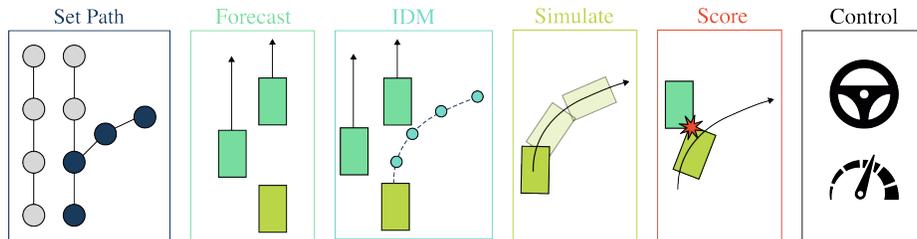


Fig. 11: PDM-Lite. Our new planner which solves all 38 scenarios of CARLA Leaderboard 2.0 has 6 stages, detailed in Section C.2.

Table 9: PDM-Lite on CARLA Official Validation Routes. We show the driving performance for each route as well as the average and their standard deviations for a total of 3 seeds. Routes 0-9 are identical to routes 10-19, except for the weather (which is ignored by PDM-Lite) and traffic behaviors. Routes 3 and 13 always crash in our setup, hence we assume the lowest possible scores.

Route	Driving Score	Route Completion	Infraction Score
0	34.9 \pm 26.3	100.0 \pm 0.0	0.35 \pm 0.26
1	51.7 \pm 14.4	100.0 \pm 0.0	0.52 \pm 0.14
2	60.5 \pm 39.2	100.0 \pm 0.0	0.61 \pm 0.39
3	0.0 \pm 0.0	0.0 \pm 0.0	0.00 \pm 0.00
4	54.4 \pm 24.2	100.0 \pm 0.0	0.54 \pm 0.24
5	35.3 \pm 11.1	68.8 \pm 0.0	0.51 \pm 0.16
6	100.0 \pm 0.0	100.0 \pm 0.0	1.00 \pm 0.00
7	9.6 \pm 1.8	85.9 \pm 0.1	0.11 \pm 0.02
8	52.5 \pm 30.2	100.0 \pm 0.0	0.53 \pm 0.30
9	32.0 \pm 24.2	100.0 \pm 0.0	0.32 \pm 0.24
10	18.3 \pm 20.9	100.0 \pm 0.0	0.18 \pm 0.21
11	25.4 \pm 30.1	100.0 \pm 0.0	0.25 \pm 0.30
12	79.6 \pm 17.7	100.0 \pm 0.0	0.80 \pm 0.18
13	0.0 \pm 0.0	0.0 \pm 0.0	0.00 \pm 0.00
14	68.0 \pm 0.4	100.0 \pm 0.0	0.68 \pm 0.00
15	31.5 \pm 15.6	68.8 \pm 0.0	0.46 \pm 0.23
16	73.3 \pm 23.1	100.0 \pm 0.0	0.73 \pm 0.23
17	7.2 \pm 9.2	95.3 \pm 8.2	0.08 \pm 0.11
18	52.5 \pm 30.2	100.0 \pm 0.0	0.53 \pm 0.30
19	100.0 \pm 0.0	100.0 \pm 0.0	1.00 \pm 0.00
Avg.	44.3 \pm 0.7	85.9 \pm 0.4	0.46 \pm 0.01

town, while Town 13 is reserved for evaluation. Each town covers an area of 10 x 10 square kilometers, encompassing varied environments such as rural, residential, and urban landscapes to replicate real-world driving conditions. The CARLA team provides a total of 90 training routes, spanning 780.6 kilometers, and 20 evaluation routes, measuring 247.6 kilometers. Every route incorporates multiple driving scenarios. We segmented these routes into shorter segments, approximately 150 meters in length and filter routes that start and end at the same position. The traffic manager within Leaderboard 2.0 initializes random background traffic around the ego vehicle comprised exclusively of 'car' entities. To enrich the dataset with greater diversity, we introduce additional vehicle classes including 'trucks', 'vans', 'bicycles', and 'motorcycles'. Moreover, we implemented randomized weather configurations for each training and evaluation route to mimic realistic driving conditions. However, night-time settings were excluded from our study due to the inadequate illumination in certain map regions. Low-light conditions significantly impede the correctness of the automatic labeling process since it is hard to obtain information about the visibility of certain objects in the image.

Data collection. We execute the expert on each of the routes and gather a comprehensive set of sensor data. The sensor data includes: (1) *RGB image*, (2) *LiDAR point cloud*, (3) *semantic segmentation images*, (4) *depth maps*, (5) *Bird's Eye View (BEV) semantic segmentation*. While DriveLM-Agent leverages only RGB images, retraining TransFuser++ needs the additional data for the auxiliary tasks. In addition, we extract privileged information from the simulator about the status of the static and dynamic objects in the scene, as follows:

- *Ego vehicle*: 3D bounding box, speed, brake, id
- *Other vehicles*: 3D bounding box, number of lidar points inside BB, distance to ego, speed, steer, throttle, brake, id, color, vehicle type, number of wheels, traffic light state, lane information (i.e., on which road and lane is the vehicle driving), vehicle in junction or not, distance to next junction, next high-level command
- *Pedestrians*: 3D bounding box, number of lidar points inside BB, gender, age, distance to ego, speed, id, lane information
- *Traffic lights*: 3D bounding box, distance to ego, state, affects ego vehicle
- *Stop signs*: 3D bounding box, distance to ego, affects ego vehicle
- *Static cars (parked cars)*: 3D bounding box, lane information
- *Landmarks (e.g., speed signs)*: 3D bounding box, distance to ego, id, text, value
- *Weather*: weather parameters

Language labels. Based on the information we extract from the simulator we create questions and answers with hand-crafted sentence templates. For more linguistic diversity and to prevent overfitting to those sentence structures those sentences could be further augmented with current state-of-the-art language models like GPT-4. However, in this work, we use a version of the dataset that is not augmented.

D DriveLM-Metrics

In this section, we offer a detailed introduction to DriveLM-Metrics. DriveLM-Metrics can be broadly categorized into three parts: P_{1-3} VQA Metrics, Behavior Task Metrics, and Motion Task Metrics.

D.1 P_{1-3} VQA Metrics

We assess the performance of P_{1-3} using common VQA metrics, and we introduce the *GPT score* for a more semantically comprehensive evaluation of our QA results. Additionally, given the graph structure of our QA, we propose the *Completeness* score to provide a thorough assessment.

BLEU [64] measures the similarity between a generated text and one or more reference texts. It operates by comparing n-grams in the generated text to those in the reference texts, with higher precision indicating a better match. However, the BLEU score exhibits insensitivity to semantic nuances and variations in word order.

ROUGE_L [52] calculates scores with the longest common sub-sequence of the model outputs and the reference answers. Similar to the BLEU metric, ROUGE is used to assess the level of matching between generated results and standard references, with the key difference being that ROUGE is based on recall.

METEOR [45] takes into account precision, recall, stemming, synonymy, stemming, and word order. It establishes alignment between model outputs and references, computes the 1-gram matching between them, and then applies penalties based on chunk blocks, providing a more nuanced evaluation.

CIDEr [86] combines elements from BLEU and vector space models. The underlying concept involves treating each sentence as a document, calculating its n-gram TF-IDF vector, and using cosine similarity to measure the semantic consistency between candidate and reference sentences. CIDEr captures matches between n-grams of different lengths and differentiates the importance of various n-grams through TF-IDF weighting.

SPICE [2] first parses the text into a syntactic dependency tree using Probabilistic Context-Free Grammar [38], then maps the dependency tree into a scene graph in a rule-based manner. The scene graph describes the objects, attributes, and their relationship in the original text, and the SPICE score is computed as the F-score of the generated scene graphs from prediction and ground truth.

GPT Score is a metric provided by ChatGPT. Traditional metrics mainly assess word-level performance and may not capture semantic nuances, potentially yielding unexpected evaluation outcomes. Leveraging ChatGPT’s robust reasoning capabilities, we employ it to gauge prediction quality and derive a more rational score. ChatGPT is prompted to assign a numerical score between 0 and

100, with higher scores indicative of enhanced prediction accuracy. The detailed prompt for GPT score evaluation is shown in Table 10.

```
Messages = [
  { "role": "system", "content": f""" An evaluator who rates my answer based on the correct answer. """ },
  { "role": "user", "content": f""" Rate my answer based on the correct answer out of 100, with higher scores indicating that the answer is closer to the correct answer, and you should be accurate to single digits like 62, 78, 41, etc. This is the correct answer: {GT}. This is my answer: {Pred}. """}]
```

Table 10: Prompt for GPT score. This differs from the prompt used in DriveGPT4 [102], but the resulting score is similar.

Completeness provides a score that accounts for how many ground truth questions are correctly answered associated with a frame. For each QA, if the score of the predicted answer is above a threshold, then this QA is considered “correctly answered” and is a correct prediction, otherwise an incorrect prediction. We then calculate the accuracy, which is the ratio of the number of correct predictions to the total number of predictions. In our setting, we utilize the SPICE score and set the threshold at 0.5.

D.2 Behavior Task Metrics

We evaluate behavior predictions by classification accuracy, along with a breakdown of the overall accuracy into its steering and speed components.

Classification Accuracy is the metric we use to evaluate Behavior Prediction Task, comprising *accuracy of behavior*, *behavior speed*, and *behavior steer*. Specifically, the ground truth of the ego vehicle future trajectory is a set of N points with coordinates (x, y) under the bird’s-eye-view, noted as $\{(x_0, y_0), (x_1, y_1), \dots, (x_N, y_N)\}$. Each point denotes the offset of the future position to the current position by a number of fixed interval times. Then, the distance for x, y at each time interval is independently computed as:

$$\{x, y\}_{dist} = ((\{x, y\}_1 - \{x, y\}_0), \dots, (\{x, y\}_N - \{x, y\}_{N-1})) \quad (2)$$

The mean of x_{dist} and y_{dist} are mapped to one of the predefined bins, where each bin corresponds to a category in either speed or steering, noted as B_{speed} and B_{steer} respectively. Finally, the speed and steering categories for this trajectory form the behavior category as (B_{speed}, B_{steer}) . We compare them with the behaviors of our DriveLM-Agent outputs and calculate the related accuracy.

D.3 Motion Task Metrics

For measuring the performance of the motion stage, we use standard metrics from the nuScenes and Waymo benchmarks: average and final displacement error, (ADE, FDE), and the collision rate of the predicted trajectory.

ADE stands for Average Displacement Error, indicating the average L2 distance between the predicted trajectory and the ground truth trajectory over all predicted horizons. It is the average of the errors at 1st, 2nd and 3rd second.

FDE stands for Final Displacement Error, which measures the Euclidean distance between the predicted endpoint and the true endpoint at the last predicted step (the 3rd second).

Collision Rate accounts for the ratio of how many test frames the predicted trajectory collides with objects in over all test frames. The number reported in Table 2 of the main paper is the average of the collision rate at 1st, 2nd and 3rd second.

Note that the calculation of **ADE**, **FDE** and **Collision Rate** follows the setting used in UniAD [32] but not ST-P3 [31]. For example, in terms of the FDE and the collision rate at 3rd second, the UniAD setting will consider the error/collision rate at only this timestep, while the ST-P3 setting will consider the error/collision rate as an average over 0.5, 1, 1.5, 2, 2.5, 3 seconds. For more details, please refer to the UniAD repo discussion. Additionally, please note that errors reported on the full nuScenes validation dataset (in prior work) is not directly comparable to results reported on the DriveLM-nuScenes val split, a challenging subset of this consisting of only keyframes with intention changes.

E DriveLM-Agent

In this section, we introduce the details of DriveLM-Agent, including the graph prompting scheme and the trajectory tokenization process.

E.1 Prompting with Context

In terms of the implementation, the content of context differs during the training and inference of DriveLM-Agent, following the teacher-forcing setting [44,82] generally adopted in recurrent networks. During training, for each edge $e \in E$ in the frame, we pick the child QA. The child questions in the edges are appended with the ground truth parent QA as the context. All QA pairs are used during training, including those without context. The objective used is next token prediction, the standard approach for language modeling. During inference, the model is applied interactively in multiple rounds to get the required context predictions as inputs for each child question. Specifically, the model is prompted with the five stages of questions in the sequential order of P_1, P_2, P_3, B, M . In this order, the model can only infer the questions in the succeeding stages after getting the predicted answer from the preceding stages.

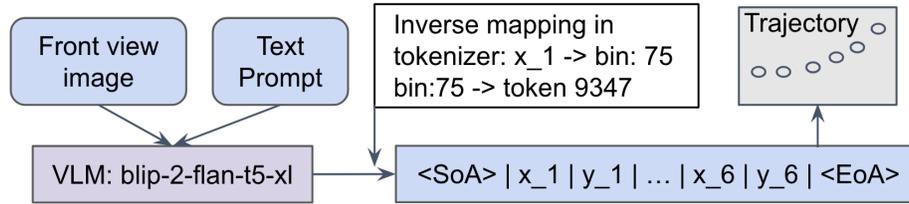


Fig. 12: Detailed architecture of DriveLM-Agent. An inverse mapping is designed to embed the real-world ego-vehicle trajectory into the token space of blip-2-flan-t5-xl.

E.2 Trajectory Tokenization Details

To generate action sequences (*i.e.*, ego future trajectories) directly with the language model we use for building the graph, we adopt the method of RT-2 [108]. This process entails the discretization and tokenization of the continuous trajectory.

Initially, we analyze the distribution of the future trajectories within the nuScenes dataset. To effectively convert the continuous (x, y) coordinate space into a discrete set of actions, we partition each coordinate axis into 256 discrete intervals. This granularity ensures a sufficient level of detail while maintaining a manageable number of tokens for the language model.

Each discretized bin corresponds to a unique token within the vocabulary of the language model. We extract the token identifiers (IDs) for numeric tokens within the vocabulary. To ensure coherence and preserve the ability to express numerical values, we omit single-digit tokens from this mapping process. Out of the remaining numeric tokens, a subset of 256 token IDs is selected to represent the trajectory data. In addition to these, we introduce two special tokens designated for marking the start and end of a trajectory sequence – the start-of-trajectory (SOT) token and the end-of-trajectory (EOT) token, respectively. This tokenization scheme enables us to encode complex trajectory information as a sequence of tokens that a language model can process. Using this mapped vocabulary, the language model can generate predicted future trajectory sequences by outputting a series of tokens, which are then translated back into the coordinate space.

F Experiments

In this section, we introduce the details of experiments, including implementation details of each subsection in **Section 4** in the main paper, more metrics on the VQA part, more ablation and comparison on computational complexity.

F.1 Implementation Details

Here we explain the implementation details for the training and validation settings used in our experiments of each of each subsections in **Section 4** in the main paper.

Fine-tuning Details. We configure the learning rate as 0.0001, no learning-rate scheduler, random seed as 1234, and other settings following the default LoRA [30] configuration. For the BLIP-2 model, we use a maximal sequence length of 400, and other hyperparameters remain the same as the official BLIP-2 implementation.

Implementation Details for Experiment in Section 4.1 & 4.3 & 4.4.

During training, we utilize all QAs as input per frame, with a subset of them having contexts (questions from $P_{2,3}$, B , and M). The contexts are extracted from ground truth, following the teacher-forcing setting [44,82] generally adopted in recurrent networks. As for inference, due to the variant complexity of the scenarios, the count of P_{1-3} QA per frame is highly imbalanced across the dataset, with a variance of over 260 on DriveLM-nuScenes. To balance the impact of this, we compute the GVQA Scores on only a subset of QA associated with each frame. To extract the QA subset, we design a set of QA patterns for each stage based on the questions generally associated with that stage. We ensure that for all our validation frames, each stage has at least one question matched with the designed pattern. In this process, except for the questions in stage P_1 , all questions in other stages have context from the previous stage’s QA, where the answers are derived from the prediction in the preceding steps. Two specific graph structure examples can be found in Fig. 13 and Fig. 14.

Implementation Details for Experiment in Section 4.2. The model is trained in the same scheme as in Section 4.1 & 4.3 & 4.4, and the training set is the DriveLM-nuScenes train split. During the inference, as there are no annotation on the Waymo (not even questions), we devise the question in a rule-based manner. Specifically, we re-use the general-purpose questions in the perception stage from DriveLM-nuScenes for Waymo as the starting questions. Then we try to find if there is any objects in the answer that is matched in the DriveLM-nuScenes annotation, such as “pedestrians”, “cars”, “trucks” and so on. Then, we generate the questions based on those matched objects automatically, which serve as the following questions in the prediction and the planning stages. A specific graph structure example can be found in Fig. 15.

Implementation Details for Experiment in Section F.2. For the generalization experiment in Section F.2 we add two new questions to the graph: (1) *Is there a person in the scene?* and (2) if the answer to the first question is yes we ask *What should the vehicle do based on the pedestrian that is crossing the road?* if the answer is no we ask *What should the ego vehicle do?*. The answer to the second question gets concatenated to the context of the final behavior question. Three examples can be found in Fig. 16.

Table 11: Generalization in DriveLM-CARLA. All methods underperform on DriveLM-CARLA-ped with the novel pedestrian object, but DriveLM-Agent can be significantly improved by including a pedestrian-specific question in its GVQA graph. move this to supp??

Method	DriveLM-CARLA (B)			DriveLM-CARLA-ped (B)		
	Acc. \uparrow	Spd. \uparrow	Str. \uparrow	Acc. \uparrow	Spd. \uparrow	Str. \uparrow
TransFuser++ [36]	70.19	73.29	90.68	8.72	8.72	100.00
DriveLM-Agent	59.63	61.50	78.26	4.59	4.59	100.00
+ Pedestrian QA	52.17	55.28	77.64	27.04	27.04	100.00
<i>DriveLM-Agent (GT)</i>	<i>60.25</i>	<i>65.22</i>	<i>80.12</i>	<i>20.92</i>	<i>20.92</i>	<i>100.00</i>
<i>+ Pedestrian QA</i>	<i>60.25</i>	<i>65.22</i>	<i>80.12</i>	<i>92.35</i>	<i>92.35</i>	<i>100.00</i>

F.2 Generalization to Unseen Objects

Next, we evaluate zero-shot generalization to novel objects. DriveLM-CARLA is collected without any pedestrians in the training or validation splits. We now generate a new test set called DriveLM-CARLA-ped, which only consists of frames where a pedestrian is present in the scene. The correct behavior is to stop for the pedestrian.

Baselines. For this experiment, we compare DriveLM-Agent to TransFuser++ [36], the state-of-the-art for CARLA. It uses a larger input image, an additional LiDAR sensor, and several driving-specific annotations (depth, semantics, 3D bounding boxes, HD map) in comparison to DriveLM-Agent. However, because of these task-specific inputs and outputs, TransFuser++ can only be trained on the base DriveLM-CARLA dataset and cannot incorporate general computer vision data during training, which makes generalization more challenging.

DriveLM-Agent. Taking advantage of the more general architecture of a VLM, we include samples from COCO [53] and GQA [34] along with DriveLM-CARLA during training for DriveLM-Agent. We compare several versions: (1) we investigate the addition of a new P_1 question during inference, ‘*Is there a person crossing the road?*’ (+ Pedestrian QA). (2) As an upper bound, we directly input the ground truth P_{1-3} graph to the model during inference, instead of the model’s predictions. For more details, please refer to the supplementary material.

Results. We present our findings in Table 11. We observe that TransFuser++ struggles on DriveLM-CARLA-ped relative to DriveLM-CARLA, with a drop in accuracy from 70.19 to just 8.72. DriveLM-Agent experiences a similar drop from 59.63 to 4.59. However, adding the pedestrian QA significantly boosts performance on the generalization setting to 27.04, albeit with slightly reduced accuracy on regular scenes. This is mainly attributed to the fact that the VLM is not able to detect all pedestrians correctly. This indicates that the large performance gains of recently published VLMs [26, 65] can support even better generalization ability in the domain of driving. Additionally, when the pedestrian QA is provided in the privileged setting that assumes access to perfect context for each question in the graph, DriveLM-Agent achieves a near-perfect score

Table 12: Graph-structured reasoning facilitates improved VQA with VLMs. Completeness measures how many questions are correctly answered in one frame of data. The improvement trends are not consistent across different conventional metrics, thus we need the GPT score as the main metrics as it evaluates the performance more comprehensively.

Context	DriveLM-nuScenes													
	Off-the-Shelf BLIP-2						DriveLM-Agent							
	BLEU-4 ↑	METEOR ↑	CIDEr ↑	ROUGE_L ↑	SPICE ↑	GPT ↑	Comp. ↑	BLEU-4 ↑	METEOR ↑	CIDEr ↑	ROUGE_L ↑	SPICE ↑	GPT ↑	Comp. ↑
None	0.022	3.317	0.1185	7.205	4.336	42.97	1.064	51.89	35.81	2.463	66.79	42.56	71.39	30.04
Graph	0.022	3.882	0.0771	7.397	7.710	45.21	0.859	53.09	36.19	2.786	65.58	49.54	72.51	31.66
<i>GT</i>	<i>0.022</i>	<i>4.397</i>	<i>0.0758</i>	<i>8.033</i>	<i>8.192</i>	<i>41.10</i>	<i>1.315</i>	<i>53.06</i>	<i>36.64</i>	<i>3.069</i>	<i>66.69</i>	<i>50.29</i>	<i>72.94</i>	<i>32.41</i>

(20.92 → 92.35) on the frames with pedestrians. Note that DriveLM-CARLAped only contains pedestrians crossing a straight road, so all models obtain a 100% accuracy on the steering class (which is always `straight`).

F.3 Results with More Metrics in VQA

In Table 12, we provide the performance under BLEU-4 [64], METEOR [45], CIDEr [86] and ROUGE-L [52] of the Table 4 in the main paper. One key observation is that different metrics reflect different trends in the performance, and the improvement is not consistent across all these metrics. This brings us the motivation to use the GPT Score as the main metric in the VQA evaluation part.

F.4 Stage-wise Ablation on Zero-shot Generalization across Sensor Configurations

In Table 14, we provide more settings of context in the zero-shot generalization across sensor configurations as in Table 2 in the main paper. One key observation is that the higher the accuracy of the behavior task, the better the performance of the motion task. With more context in the behavior task, the accuracy improvement mainly originates from the improvement of the speeding accuracy, which finally affects the FDE score.

F.5 More VLMs evaluated on DriveLM-nuScenes

Table 13: More VLMs evaluation on DriveLM-nuScenes. On the base model of LLaMA-Adapter V2, we observe a slight improvement by using multi-frame as input.

Table A. Base VLM	Behavior Context	Motion Context	Behavior (B)			Motion (M)	
			Acc. ↑	Speed ↑	Steer ↑	ADE ↓	Col. ↓
blip-2-flan-t5-xl	Graph	B	57.49	69.89	80.63	1.74	1.89
LLaMA-Adapter V2	Graph	B	55.31	61.97	81.43	1.86	2.03
LLaMA-Adapter V2*	Graph	B	57.83	67.21	83.49	1.75	1.69

Table 14: Zero-shot Generalization across Sensor Configurations. B^* denotes using ground truth behavior QA as context for motion task. Results on 1k randomly sampled frames from the Waymo val set after training on DriveLM-nuScenes. Our key observation is that the higher the accuracy of the behavior task, the better the performance of the motion task.

Method	Behavior	Motion	Behavior (B)			Motion (M)	
	Context	Context	Acc. \uparrow	Speed \uparrow	Steer \uparrow	ADE \downarrow	FDE \downarrow
Command Mean	-	-	-	-	-	7.98	11.41
UniAD-Single	-	-	-	-	-	4.16	9.31
BLIP-RT-2	-	-	-	-	-	2.78	6.47
DriveLM-Agent	None	B	35.70	43.90	65.20	2.76	6.59
	P_1	B	38.20	43.67	70.74	2.67	6.41
	P_2	B	39.52	44.20	78.67	2.62	6.19
	P_3	B	34.62	41.28	64.55	2.85	6.89
	P_{1-3}	B	39.73	54.29	70.35	2.63	6.17
<i>BLIP-RT-2</i>	-	B^*	<i>100.0</i>	<i>100.0</i>	<i>100.0</i>	<i>2.41</i>	<i>5.79</i>

In terms of validating the generality of the proposed DriveLM-Agent and exploring if the video input helps the model’s performance in this task, we select LLaMA-Adapter V2 (7B) [1] as it is compatible with single- and multi-frame inputs. The results in Table 13 (within 9 training epochs) show performance similar to BLIP-2.

Note that due to the large variety of question types in DriveLM-nuScenes, analyzing each question would be prohibitively expensive.

Instead, we provided an ablation of the effectiveness of each stage (P_{1-3}) in Table 14, and a representative-question-wise analysis.

We have to admit that it is non-trivial in improving the QA performance (P_{1-3}) under current question setting and we found that it had little affect on the final task performance (behavior and motion).

This inspired us that maybe the question setup worth more effort to explore, and with well-set question, the model could be “prompted correctly” to understand the scene and perform the downstream task better.

F.6 Computational Complexity

In Table 5 of the main paper, we provide a comparison of the computational complexity of DriveLM-Agent to UniAD-Single. A future direction would be caching the vision tokens and batching the different question patterns, which can speed up the inference time fundamentally.

G Qualitative Results

In this section, we show the qualitative results of the experiments, including VQA on our DriveLM-nuScenes, generalization results across sensor data on Waymo, and generalization results to unseen objects on DriveLM-CARLA.

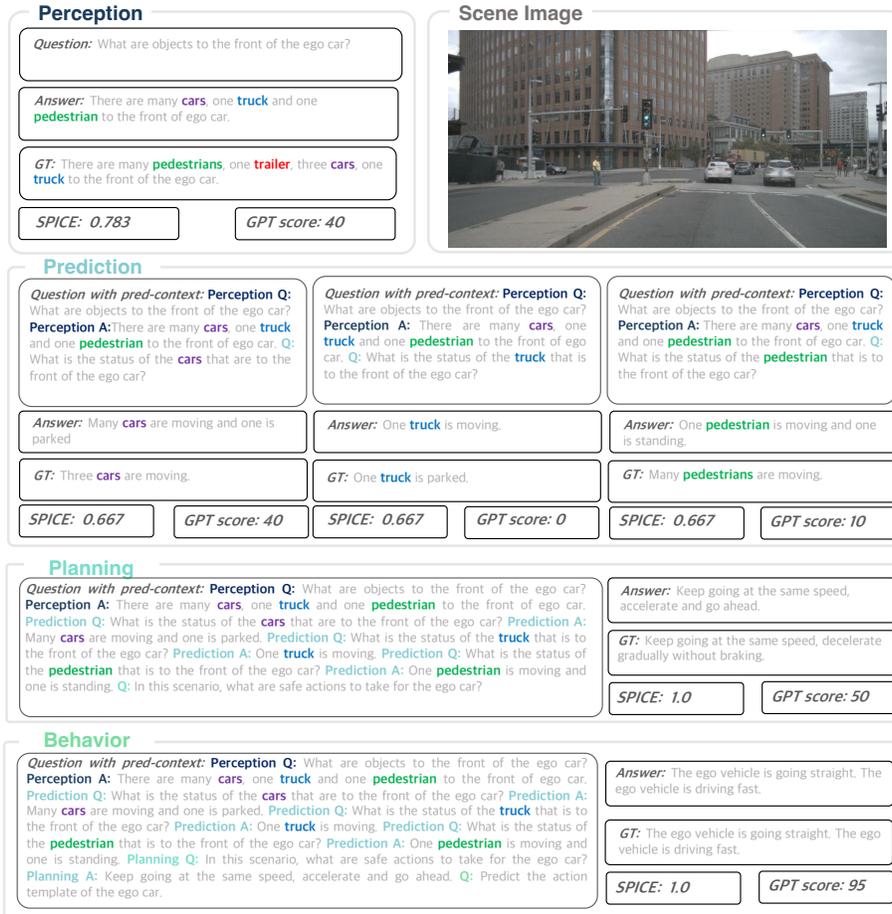


Fig. 13: Detailed qualitative results on DriveLM-nuScenes. The graph prompting process can be divided into different tasks, and different QAs in each task revolve around different objects.

G.1 DriveLM-nuScenes

This section shows the qualitative examples for the DriveLM-nuScenes. In Fig. 13, we showcase a detailed example of GVQA reasoning process on DriveLM-nuScenes, encompassing P_{1-3} QA and the behavior task. We compare the predicted answers with ground truth and provide SPICE scores and GPT scores. In this figure, the second question in the prediction stage represents a typical error. Due to the input of single-frame images, our model often struggles to accurately determine the correct movement status of objects. This judgment is indeed challenging even for humans. Furthermore, in Fig. 14, we present additional qualitative results to showcase our model’s performance.

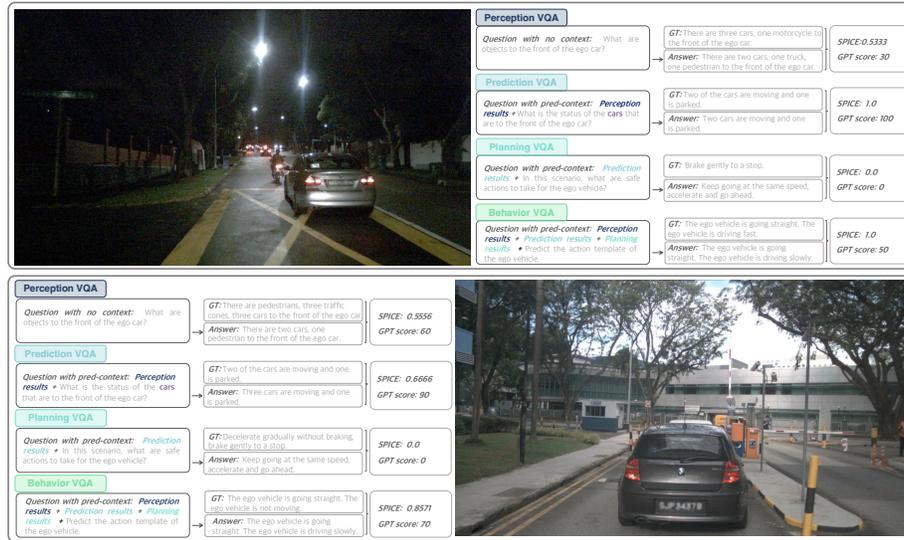


Fig. 14: More qualitative results on DriveLM-nuScenes. The examples in the figure illustrate the robust ability of our DriveLM-Agent to perform VQA tasks in driving scenarios.

G.2 Waymo

This section demonstrates the generalizability of our model across sensor configurations. Fig. 15 illustrates the results of our model, trained on DriveLM-nuScenes, when applied to inference on Waymo. As we do not annotate data on Waymo, the questions are manually defined, and ground truth is not provided. These results showcase the robust generalization capability of our model.

G.3 DriveLM-CARLA

In this section, we provide qualitative examples for the CARLA dataset.

Generalization to the unseen pedestrian scenario. Fig. 16 shows the generated behaviors for the generalization test set on the unseen pedestrian scenario. The first example, illustrated on the top of Fig. 16, demonstrates a successful situation where DriveLM-Agent accurately recognizes a pedestrian. It subsequently infers the appropriate action to undertake, which in this case, is to stop the vehicle. The behavior generation is able to interpret this context, resulting in the correct behavior pattern, as evidenced by the ego vehicle coming to a complete stop. The other two examples represent the predominant failure modes of DriveLM-Agent in scenarios involving pedestrians. The middle example of Fig. 16 shows the case where the model still detects the pedestrian. However, it fails to translate this detection into the correct behavior. The final example, shown at the bottom, highlights a more severe limitation where the model

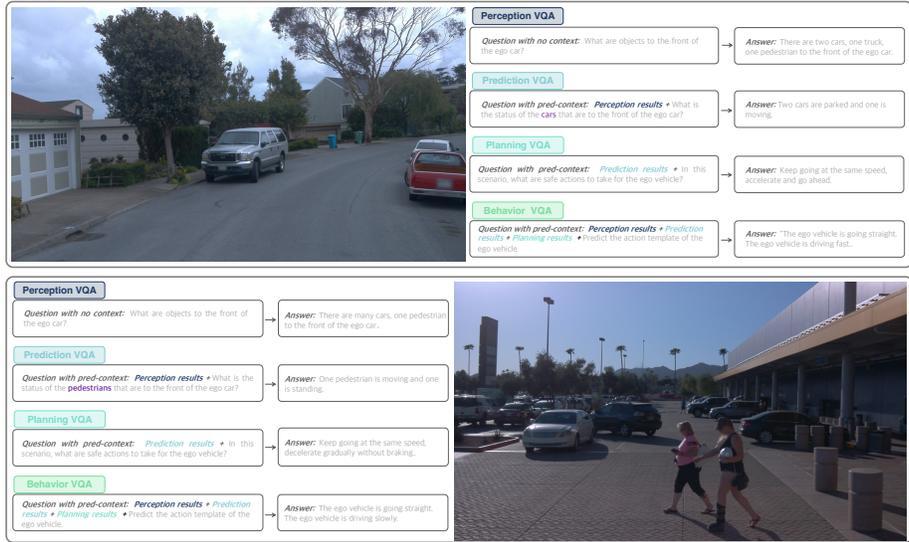


Fig. 15: Qualitative results on Waymo. We present two examples showcasing the generability of DriveLM-Agent to new sensor configurations, demonstrating the strong generalization capability of DriveLM-Agent.

completely overlooks the pedestrian. In such instances, DriveLM-Agent acts as if the pedestrian is non-existent, which consequently results in it not executing any evasive or stopping maneuvers, posing a significant risk in a real-world scenario.

Graph Visual Question Answering. This section presents two examples of the graph visual question answering tasks using the CARLA dataset to evaluate the performance of DriveLM-Agent (Fig. 17). We only show a subset of the evaluated questions. In the first example the ego vehicle drives behind another vehicle. The primary task is to follow the road and adjust the speed in accordance with the leading vehicle. Our results indicate that DriveLM-Agent demonstrates proficient scene understanding by accurately identifying all important objects in the scene. Despite the ground truth data indicating occasional inaccuracies in object color labeling by the CARLA simulator, DriveLM-Agent maintains reliable performance in object recognition. Additionally, the model can identify the vehicle in front and reason about what to do based on the leading vehicle.

The second example takes place at an intersection regulated by a stop sign. DriveLM-Agent identifies all objects and can reason about the situation. It correctly identified that it needs to stop not simply due to the stop sign, but primarily because of a motorcycle positioned ahead. This implies that DriveLM-Agent is capable of prioritizing dynamic obstacles over traffic control devices under certain circumstances.

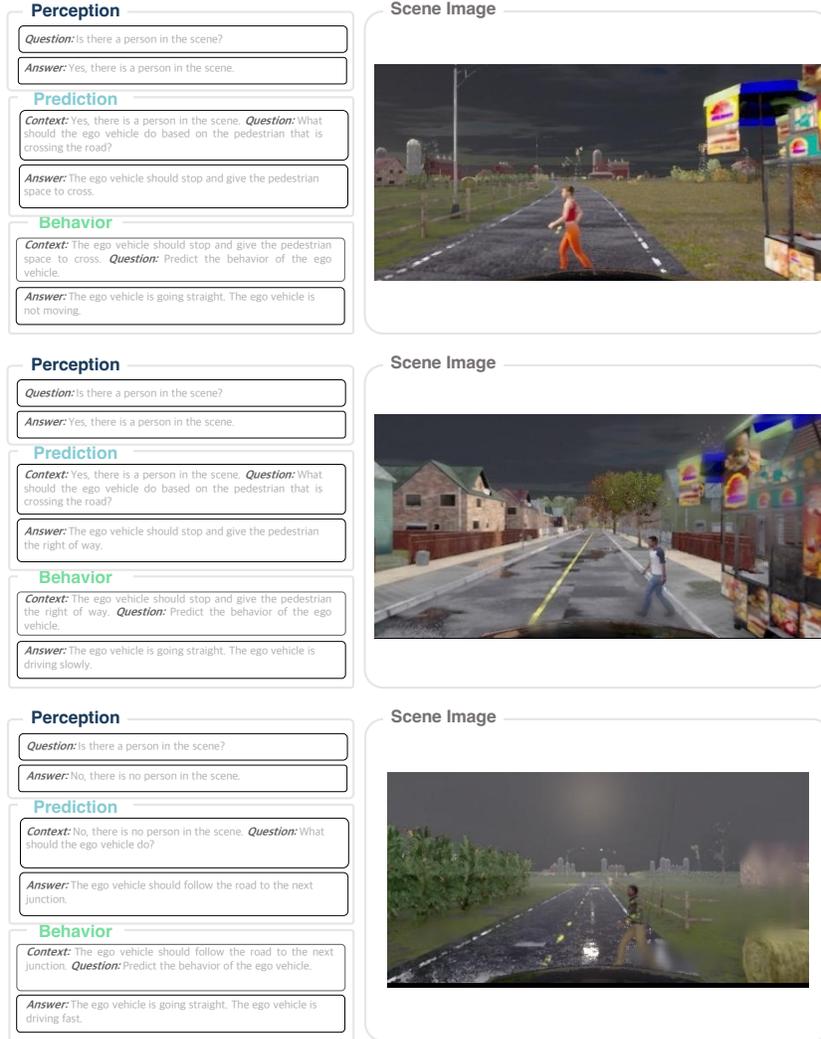


Fig.16: Qualitative results on the generalization test set in DriveLM-CARLA. We show three examples of DriveLM-Agent handling the pedestrian scenario. The first example shows a success case and the second and third show two common failure cases of the model.

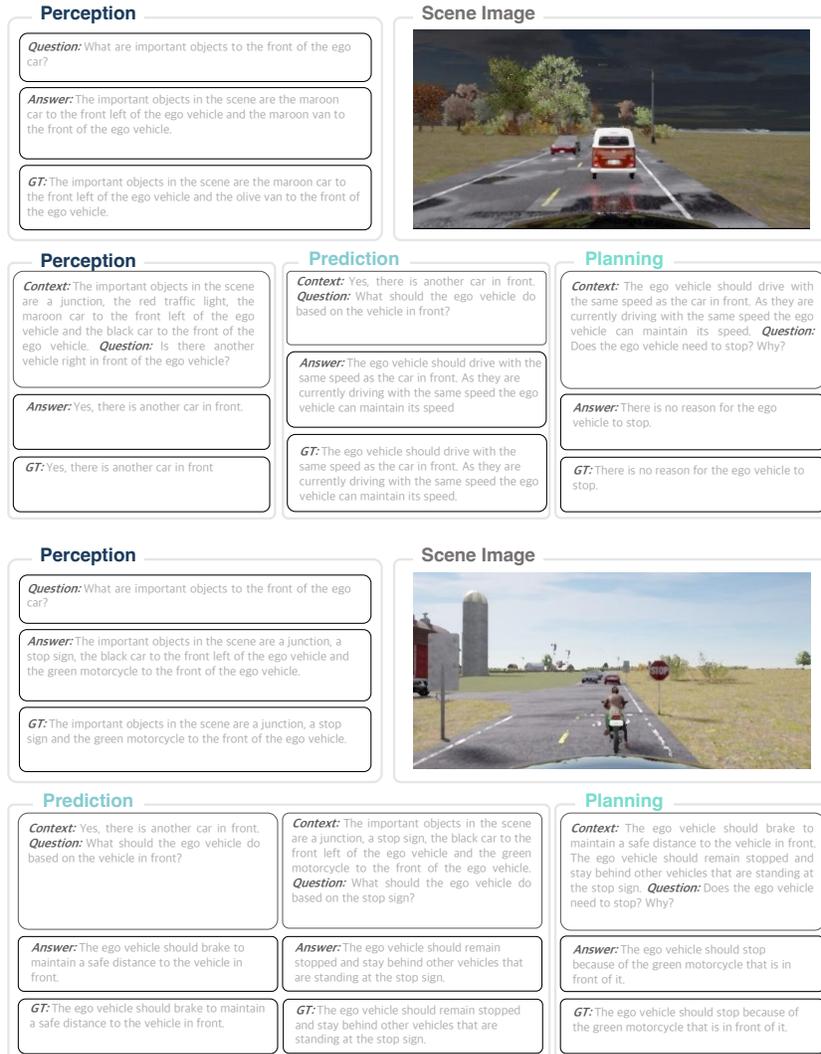


Fig. 17: Qualitative VQA results in DriveLM-CARLA. The first example shows how DriveLM-Agent deals with a situation with a vehicle directly in front of the ego vehicle. The second example shows an intersection with a stop sign and other traffic participants waiting in front of the ego vehicle at the stop sign.

H More Related Work

In this section, we supplement the related work mentioned in the main paper.

Reasoning Over Graph Structures. Reasoning is one of the basic forms of simulated human thinking, enabling the derivation of new judgments from one or several existing judgements [12]. Many prior reasoning works have been grounded in graph-based approaches [5, 13, 75, 92]. XNMs [75] employs scene graphs for explainable and explicit reasoning with structured knowledge. KPRN [92] utilizes knowledge graph for reasoning and applies it to recommender systems. GoT [5] models LLM-generated information as an arbitrary graph and brings the LLM reasoning closer to brain mechanisms. Inspired by these successful attempts, we try to link the stages of perception, prediction, and planning in autonomous driving through a graph, enabling the model to grasp the reasoning process and deduce unseen scenarios based on learned graph structure.

Embodied Planning with LLMs Recent work endeavors to leverage the formidable reasoning and generalization capacity of LLMs [25, 43, 83] for embodied AI systems [23, 33, 35, 40, 51, 69, 108]. PaLM-E [23] trains an LLM for various embodied tasks including sequential robotic manipulation planning. CaP [51] provides a robot-centric formulation of language model generated programs executed on real systems. RT-2 [108] represents robot actions as language tokens, training vision-language models to output robot policies. These methods showcase the capabilities of LLMs in embodied planning tasks, inspiring us to apply them to address the current shortcomings in generalization in AD, which is far less explored.

Vision-language Benchmarks for Driving. An increasing number of vision-language datasets have been proposed for AD systems [20, 41, 42, 56, 59, 67, 71, 97, 98, 101, 107]. NuScenes-QA [67] and NuPrompt [98] provide perceptual information as text by describing the positions and states of surrounding objects. BDD-X [42] provides reasons for the ego vehicle’s actions in natural language descriptions. DRAMA [56] and Rank2Tell [71] identify crucial objects and provide corresponding driving suggestions. However, these datasets focus on scene-level context or individual objects. DriveLM fills this gap in the literature by organizing language annotations from object-level and task-level with a graph structure.

I Broader Impact

We believe that this approach can accelerate progress in the field of autonomous driving by enabling it to directly benefit from better VLMs. Our goal is to make progress towards autonomous driving, which will have profound impact if successful. We recognize that by bringing VLMs into this area, we accept their ethical implications, such as hallucinations and high resource use. Yet, by improving the interactivity between humans and autonomous driving systems, we can build confidence in the technology. This could hasten its acceptance and lead to safer transportation in the long term.