World Knowledge-Enhanced Reasoning Using Instruction-Guided Interactor in Autonomous Driving

Mingliang Zhai[1,2,3], Cheng Li[1,3], Zengyuan Guo[3], Ningrui Yang[1,3], Xiameng Qin[3], Sanyuan Zhao[1], Junyu Han[3], Ji Tao[3], Yuwei Wu[2,1]*, Yunde Jia[2,1]

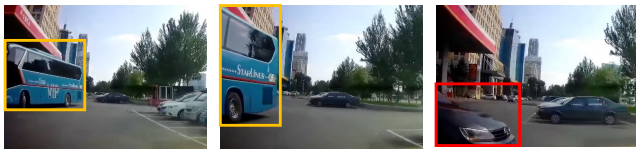
[1]Beijing Institute of Technology
[2]Shenzhen MSU-BIT University
[3]Chongqing Changan Automobile Co., Ltd.
{zhaimingliang, licheng, ningrui, zhaosanyuan, wuyuwei, jiayunde}@bit.edu.cn
{guozy, qinxm, hanjunyu, taoji}@changan.com.cn

Abstract

The Multi-modal Large Language Models (MLLMs) with extensive world knowledge have revitalized autonomous driving, particularly in reasoning tasks within perceivable regions. However, when faced with perception-limited areas (dynamic or static occlusion regions), MLLMs struggle to effectively integrate perception ability with world knowledge for reasoning. These perception-limited regions can conceal crucial safety information, especially for vulnerable road users. In this paper, we propose a framework, which aims to improve autonomous driving performance under perception-limited conditions by enhancing the integration of perception capabilities and world knowledge. Specifically, we propose a plug-and-play instruction-guided interaction module that bridges modality gaps and significantly reduces the input sequence length, allowing it to adapt effectively to multi-view video inputs. Furthermore, to better integrate world knowledge with driving-related tasks, we have collected and refined a large-scale multi-modal dataset that includes 2 million natural language QA pairs, 1.7 million grounding task data. To evaluate the model's utilization of world knowledge, we introduce an object-level risk assessment dataset comprising 200K QA pairs, where the questions necessitate multi-step reasoning leveraging world knowledge for resolution. Extensive experiments validate the effectiveness of our proposed method.

Introduction

The Multi-modal Large Models (MLLMs) alleviates the limitations of expert knowledge and training data diversity in traditional autonomous driving systems. Recent research (Wen et al. 2023; Ma et al. 2023; Tian et al. 2024b; Chen et al. 2024a; Sima et al. 2023; Cui et al. 2023; Wang et al. 2024a; Ding et al. 2024; Bai et al. 2024; Tian et al. 2024a) have made significant progress in understanding and reasoning about perceivable regions. However, there remain deficiencies in handling perception-limited regions, *e.g.*, occluded areas caused by dynamic or static obstacles such as bus and buildings. As shown in Figure 1, autonomous driving systems typically plan and control only within the perceived areas, while hidden potential risks are critical factors leading to severe accidents. These occluded areas may conceal information crucial to road safety, especially for undetected vulnerable road users, such as pedestrians and cyclists, who are particularly susceptible to the effects of these occlusions. We consider that a promising solution is to utilize instruction-guided extraction of highly aggregated visual embeddings to fully leverage the world knowledge encoded in multi-modal large language models for inference.



(a) Risk caused by occlusion of moving objects

(b) Risks present in static environments

Figure 1: **Examples of dynamic and static environments risks.** (a) The bus in motion severely obstructs the line of sight, resulting in the black sedan being hidden, which significantly increases the risk of a traffic accident in an unprotected scenario. (b) Buildings in static scenes can also become occluding objects. For example, in a construction site scene, the construction gate blocks the workers behind the gate.

Currently, methods utilizing MLLMs for driving tasks are primarily categorized into the following three types: 1. Fine-tuning MLLMs (Wang et al. 2024a; Ding et al. 2024; Wen et al. 2023; Sima et al. 2023; Cui et al. 2023; Fu et al. 2024) directly for tasks such as prediction and planning. 2. The dual-branch system (Tian et al. 2024b; Ding et al. 2023; Mei et al. 2024) for separating and managing tasks based on real-time requirements, addressing time constraints with fast and slow branches. 3. The training-free method (Dewangan et al. 2023; Wang et al. 2024b; Ma et al. 2023) based on the chain of thought. These three types of methods have shown promising results, but there are two main

*Corresponding author: Yuwei Wu

issues. Firstly, MLLMs are not well-suited for multi-view video inputs, which limits the model's ability to fully leverage perception ability and integrate world knowledge into subsequent reasoning processes. Secondly, due to the constraints on the input sequence length of MLLMs, aligning inputs with widely used autonomous driving systems is challenging.

In this paper, we propose a multi-view multi-modal unified architecture, which aim to integrate perception ability and world knowledge. The core of the architecture is the instruction-guided interaction module to adapt multi-view video inputs and enhancing the correlation between visual features and natural language instructions, facilitating pre-fusion of features across views and modalities. We select the top-k most similar visual features as visual queries, integrating these queries with original visual features using a cross-attention mechanism to generate enhanced and highly aggregated visual representations. This pre-fusion strategy not only aids subsequent decoders in more efficient inference but also significantly reduces the length of input sequences, thereby adapt to the inputs of autonomous driving systems.

To align between multi-view video feature and language embedding space, we collected and refined a large-scale visual-textual dataset aimed at supporting highly complex scene understanding and response capabilities. This dataset comprises over 1.7 million annotated location entries and 2 million dialogue records, covering a diverse range of real-world scenarios. Furthermore, to address specific corner cases, we employ GPT-4o (for multi-modal information extraction) and GPT-4o-mini (for pure text reasoning path generation), selecting challenging scenarios from NuScenes such as occlusions, traffic violations, and potential collision risks. For these scenarios, we conduct thorough object-level risk assessments. Based on these efforts, we have design a dataset of 200K QA pairs for training a deeper understanding of complex scenes and to evaluate reasoning abilities in perception-limited regions.

In summary, our approach aims to leverage instruction-guided visual embeddings to handle multi-view video data inputs, enhancing the integration of perception ability and world knowledge, and achieving autonomous driving under constrained perception conditions. Our contribution can be summarized as follows:

- We propose a multi-modal large language model architecture tailored for autonomous driving systems, which enhances the perception ability of MLLMs and integrates world knowledge to enable reasoning in perception-limited regions.
- We introduce a plug-and-play instruction-guided interaction module that employs a pre-fusion strategy to generate highly aggregated visual features. This module not only facilitates more efficient inference processes in subsequent decoders but also significantly reduces the input sequence length.
- We have reorganized the existing datasets for align between multi-view video feature and language embedding space, and propose an object-level risk assessment dataset for evaluating inference performance in perception-limited scenarios.

Related Works

MLLMs with World Knowledge

Existing Large Language Models (LLMs) have demonstrated extensive world knowledge (Yu et al. 2024), which plays a crucial role in multi-hop reasoning tasks. Certain LLMs, such as GPT-4 (Achiam et al. 2023), ChatGLM2 (GLM et al. 2024), and LLaMA (Touvron et al. 2023), exhibit strong performance on knowledge-driven tasks. Recently, MLLMs have introduced world knowledge into the multi-modal domain. Some MLLMs, like CLIP (Radford et al. 2021) and ALIGN (Cohen 1997), use contrastive learning to create similar embedding spaces for language and vision. On one hand, models like LLaVa (Liu et al. 2024b), PaLM-E (Driess et al. 2023), PaLI (Chen et al. 2022), RT2 (Brohan et al. 2023), and InternVL (Chen et al. 2024b) align images and text tokens using self-attention by interweaving or concatenating tokens of fixed sequence length. On the other hand, models such as Flamingo (Alayrac et al. 2022), Qwen-VL (Bai et al. 2023), and BLIP-2 (Li et al. 2023) employ static queries for cross-attention with visual features to extract a fixed number of visual tokens. These approaches effectively map visual features into the linguistic space to leverage world knowledge for reasoning. However, the utilization of world knowledge is often language-based, and when dealing with multi-perspective video data, visual tokens dominate the input token sequence, thereby diminishing the exploitation of world knowledge. We propose a world-knowledge-enhanced MLLM architecture that aggregates visual tokens effectively and maximizes the utilization of world knowledge.

MLLMs for Driving Tasks

For driving tasks, multi-view images or videos are typically required as input. Approaches for handling multiple image inputs can be categorized into image feature fusion (Awadalla et al. 2023; Laurençon et al. 2024; Lin et al. 2024) and image concatenation (Jiang et al. 2024; Sun et al. 2024). The former approach significantly reduces the resolution of the input images, leading to a loss of image details. The latter approach substantially increases the input sequence length. Our model adopts a novel approach where relevant features are extracted based on user instructions, and potential details lost are supplemented from the original features.

Previous work typically fine-tunes existing MLLMs with driving tasks. Most existing MLLMs are optimized primarily for visual understanding. As a result, autonomous driving MLLMs fine-tuned using these models (Sima et al. 2023; Wang et al. 2023; Ma et al. 2023; Ding et al. 2024; Tian et al. 2024b,a; Bai et al. 2024) often lack fundamental 3D understanding and behavioral reasoning capabilities. Recent work (Wang et al. 2024a) has integrated detection heads into query transformer. The latent queries used for token extraction also interact with detection queries to guide the tokens to capture 3D perception information. However,

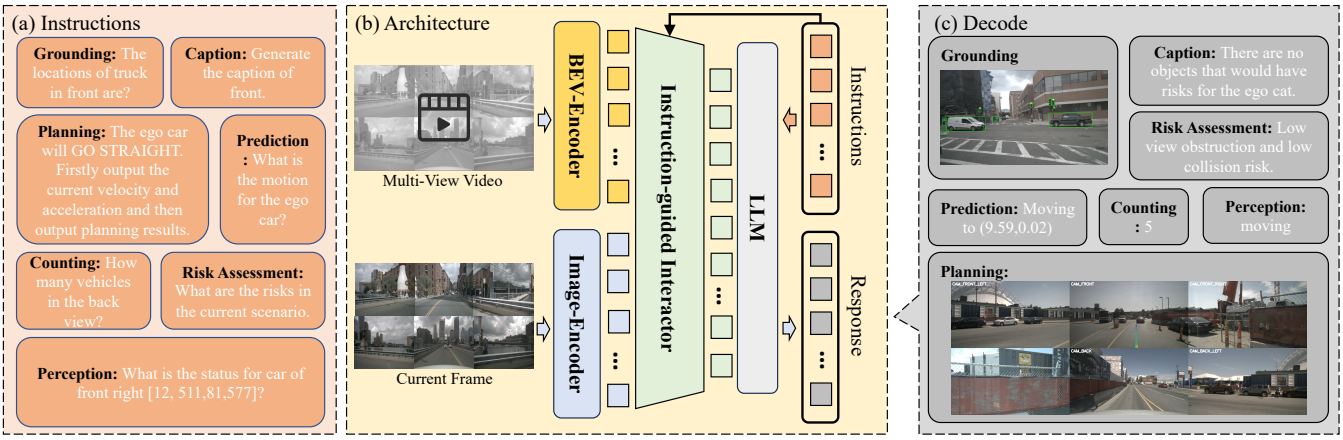


Figure 2: **Overall of our architecture.** (a) Task-specific instructions. (b) A multi-modal large language model equipped with an interactor, which can select important tokens and perform pre-fusion of these tokens before inputting multi-view and multi-modal information into the LLM. (c) Decoding results and visualization of tokens output by LLM.

for perception-limited regions, it is necessary not only to achieve comprehensive perception of the current scene but also to integrate world knowledge for reasoning.

Method

Architecture

As illustrated in Figure 2, our overall architecture comprises four key components: (1) a shared visual encoder f_{enc}, (2) a BEV encoder f_{bev}, (3) a instruction-guided interactor $f_{interact}(\cdot)$ that extracts relevant visual tokens based on user requests, and (4) a large language model (LLM) $f_{LLM}(\cdot)$ to receive visual and language instruction tokens to generate the response.

We input multi-view video sequence $\mathbf{V} = \{V^i\}_{i=0}^{N_{view}} = [v_1^i, v_2^i, v_3^i, \ldots, v_n^i]$, where N_{view} is the number of views (total 6 views), n is the number of frames. For clarity in the following, we use $\mathbf{L}_{inst} \in \mathbb{R}^{N_{inst} \times D}$ and $\mathbf{L}_{resp} \in \mathbb{R}^{N_{resp} \times D}$ to denote the language instruction tokens and response tokens respectively, where D denotes hidden size, N_{inst} and N_{resp} are numbers of tokens for the instruction and response. We first extract BEV features $\mathbf{F}_{bev} \in \mathbb{R}^{N_{bev} \times D}$ by BEV encoder f_{bev}, and current frame multi-view image features $\mathbf{F}_{mv} = \{F_{mv}^i\}_{i=0}^{N_{view}} \in \mathbb{R}^{N_{mv} \times D}$. Notably, after extracting visual features using the encoder f_{enc}, we employ an MLP to project the feature dimensions of the visual features to the feature dimension D of the language embeddings. And then we can formula our architecture as

$$\mathbf{L}_{resp} = f_{LLM}\Big(\mathbf{L}_{inst}, f_{interact}\Big((\mathbf{F}_{mv}, \mathbf{F}_{bev}), \mathbf{L}_{inst}\Big)\Big). \tag{1}$$

Instruction-guided Interactor. Current MLLMs often concatenate information from different modalities directly as input, and then utilizing the global attention mechanism in LLMs to interact with this information. However, the redundant multi-modal tokens can make it challenging for these models to identify useful information relevant to the task. Moreover, as the number of input images or modalities increases, the excessively long input sequences can lead to computational demands that are unacceptably high. This issue is particularly prominent in autonomous driving systems, which require inputs from multiple perspectives and modalities.

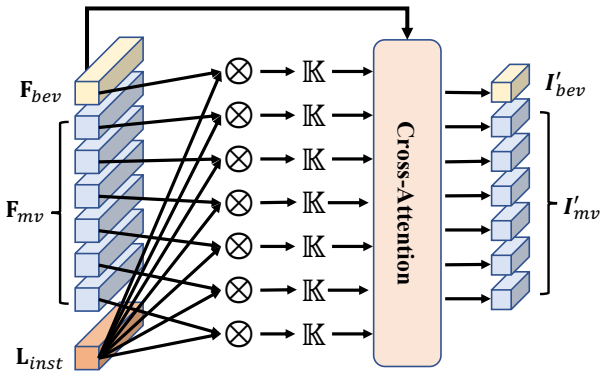


Figure 3: **Interactor Module.** $\otimes$ represents similarity operator. $\mathbb{K}$ represents the top-k operator.

To address this issue, we propose the instruction-guided interactor, which can select important tokens and pre-fuse multi-view, multi-modal information before feeding it into LLM. As shown in Figure 3, the instruction-guided interactor consists of two operations: a selection operation to identify the k tokens most relevant to the language instruction, and an interaction operation to facilitate interaction between the selected tokens and the original features. The process of the instruction-guided interactor is formulated as

$$\begin{aligned} F_{mv}^{i'} &= \mathbb{K}(F_{mv}^i \bigotimes \mathbf{F}_{inst}), \\ \mathbf{F}'_{mv} &= \{F_{mv}^{i'}\}_{i=0}^{N_{view}}, \\ \mathbf{F}'_{bev} &= \mathbb{K}(\mathbf{F}_{bev} \bigotimes \mathbf{F}_{inst}), \end{aligned} \tag{2}$$

where $\mathbb{K}$ represents the top-k operator, $\otimes$ denotes the com-

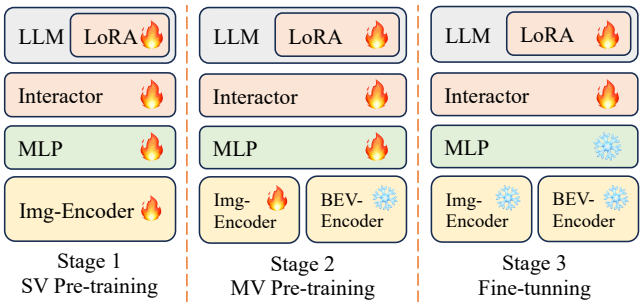


Figure 4: Training pipeline of our method. SV means single-view, and MV denotes multi-view.

putation of similarity between two matrices. Simply selecting k relevant tokens may result in the loss of some critical information. Therefore, inspired by Q-former (Li et al. 2023), we enhance the features by computing cross attention between these k tokens and the global features, which can be represented as

$$\begin{aligned} I_{mv} &= \text{CrossAttn}(\mathbf{F}'_{mv}, \mathbf{F}_{mv}, \mathbf{F}_{mv}), \\ I_{bev} &= \text{CrossAttn}(\mathbf{F}'_{bev}, \mathbf{F}_{bev}, \mathbf{F}_{bev}), \end{aligned} \tag{3}$$

where $\text{CrossAttn}(\cdot, \cdot, \cdot)$ is the standard cross-attention operation with the parameters query, key, and value, respectively, I_{mv} and I_{bev} are concatenated and fed into the LLM. Notably, the instruction-guided interactor is a plug-and-play module that can be easily extended to more modalities.

Training Strategy

Current MLLMs struggle to adapt to multi-view inputs in driving scenarios. To address this issue, we propose a three-phase training strategy. The first phase focuses on aligning the visual and linguistic feature spaces. The second phase is dedicated to constructing relationships between multi-view inputs. The third phase involves instruction fine-tuning to adapt to downstream tasks. We trained the model following the pipeline shown in Figure 4.

Stage 1: Single-view Pre-train. In this stage, we train our model on single images for captioning and grounding tasks, aiming to establish image-level, region-level, and object-level visual-language alignment. During this process, we unfreeze all parameters except LLM and utilize LoRA to train the LLM.

Stage 2: Multi-view Alignment Pre-train. To endow the MLLM with the capability to comprehend multi-view driving scenarios, we extended the dataset from the first stage to incorporate multiple views for model training and incorporated BEV features to provide global semantic information. In this phase, the trainable parameters are similar to those the first phase, and the BEV encoder is frozen.

Stage 3: Task-specific Instruction Tuning. We have integrated and cleaned multiple open-source datasets. We format all data to Llava's style and use LoRA fine-tune. After this training phase, we obtained a MLLMs capable of engaging in dialogues and exhibiting proficient performance across various driving tasks.

Task	Pairs
Grounding-NuScenes	1700k
Caption-NuScenes	100k
Total	1800k

Table 1: Details of pre-train datasets

Dataset	Train	Test
NuScenes-QA	376k	83k
NuScenes-MQA	1204k	255k
OmniDrive-NuScenes	486k	90k
NuInstruct	72k	15k
RiskAssessment	166k	35k
Total	2304k	478k

Table 2: Details of fine-tune datasets

Dataset Construction

To achieve multi-modal alignment, we collected and refined a large-scale multi perspective image text pair, including 1.7M grounding data, 200K object-level caption data (objects, risks, weather etc.), 4 open-source datasets and our object-level risk assessment dataset, total 4M samples. Then we format all the data into a unified format. Regarding the grounding data, we use a pre-trained Grounding-DINO (Liu et al. 2023b) model, specifically trained on traffic scenes, to extract all significant objects from single-view images, such as vehicles, pedestrians, traffic signs, and traffic lights.

Object-level Risks Assessment (ORA) To evaluate the model performance in perception-limited regions, we propose an object-level risks assessment dataset base on NuScenes (Caesar et al. 2019). We define four types of object-level risks: 1. View obstruction. 2. Collision possibility. 3. Traffic rule violations. 4. Potential risk. We classify the QA pairs into six categories: **Exist** determines whether there is a risk. **Level** classifies the risk into three levels—low, medium, and high. **Category** specifies one of the four risk categories mentioned earlier. **Object** identifies the category of the target causing the risk. **Reason** describes the cause of the risk. **Grounding** denotes the location of the target causing the risk.

We use GPT-4o and GPT-4o-mini to construct object-level risk assessment data. The construction process is divided into two steps: **Step 1.** We input images along with detailed object information—including category, direction relative to the vehicle, and distance from the vehicle into GPT-4o. We also specify the desired output format to obtain raw data that captures the object-level risks associated with the scene. **Step 2.** The raw data generated by GPT-4o is then processed by GPT-4o-mini. This model is used to organize the data and create diversity question-answer pairs that cover different aspects of the object-level risks identified. The specific prompts and data samples are provided in the appendix.

METHOD	Language Score				Accuracy				mAP ↑
	BLUE1 ↑	BLUE4 ↑	CIDEr ↑	ROUGE_L ↑	exist ↑	level ↑	cate ↑	object ↑	
Bunny-Llama3	62.74	39.86	244.10	59.89	58.34	77.77	69.13	71.31	0
Ours w/o Interactor	68.13	45.86	313.43	62.40	68.79	85.45	78.53	81.75	14.95
Ours w/o Selection	68.45	47.13	331.88	63.71	68.89	89.24	79.28	83.45	14.43
Full	**70.42**	**49.08**	**344.55**	**65.02**	**70.07**	**89.87**	**83.87**	**83.94**	**15.68**

Table 3: Results on ORA dataset.

METHOD	BLUE1 ↑	BLUE4 ↑	ROUGE_L ↑	ACC ↑
OPT-1.3B	**69.8**	40.4	62.6	60.4
OPT-1.3B + st	64.4	36.0	47.4	63.8
OPT-6.7B	67.4	38.4	62.4	61.1
Ours	67.4	**48.6**	**67.5**	**74.4**

Table 4: Results on NuScenes-MQA.

METHOD	CIDEr ↑	ROUGE_L ↑	METEOR ↑
OmniDrive	68.6	32.6	38.0
OmniDrive w/o online	69.0	32.7	38.2
Ours	**103.9**	**38.5**	**40.1**

Table 5: Results on OmniDrive-NuScenes.

METHOD	MAE ↓	ACC ↑	mAP ↑	BLUE4 ↑
Video-LLama	12.77	24.8	12.85	25.3
BEV-InMLLM	9.07	32.48	**20.71**	35.2
Ours w/o top-k	5.09	43.81	13.01	46.69
Ours	**4.33**	**52.71**	16.66	**69.85**

Table 6: Results on NuInstruct.

METHOD	ACC ↑
MSMDFusion+MCAN	**60.4**
CenterPoint+MCAN †	59.5
OmniDrive	59.2
Ours	58.4

Table 7: Results on NuScenes-QA. † represent use Lidar infomation.

Experiments

Implementation

We employ EVA-02-L (Fang et al. 2023) as the image encoder and a re-trained SparseBEV (Liu et al. 2023a) (excluding future frames and validation set) as the BEV encoder. For the large language model, we utilize LLaMA3-8B (Touvron et al. 2023). A 2-layer MLP with ReLU activation functions is used to map feature dimensions. In the selection operation, cosine similarity is used as the similarity metric, and a 2-layer cross-attention is employed in the interaction operation. When selecting $\mathrm{top\text{-}k}$ tokens, we set $k = 90$ for image features and $k = 300$ for BEV features.

During the single-view and multi-view alignment pre-training stage, we adopt the same strategies as LLava-Next (Liu et al. 2024a), including optimizer, learning rate, and batch size, training for 2 epochs. We use 32 Tesla A100 80G to train 3 days. For task-specific instruction tuning stage, we use the AdamW (Loshchilov and Hutter 2017) optimizer, setting the learning rate to 1×10^{-5} and a batch size of 8. To promote training stability and convergence, we implement a cosine annealing learning rate schedule with a warm-up period. For this stage, we use 8 Tesla A100 80G GPUs, and the training is conducted over a period of 8 hours.

Matrics

For caption task such as scene description and risk assessment, we employ commonly used language-based metrics to evaluate word-level sentence similarity, including BLEU, ROUGE_L, and CIDEr. Notable, for data in NuScenes-MQA (Inoue et al. 2024) with tagged parts and OmniDrive-NuScenes (Wang et al. 2024a), we compute the Accuracy metric. For grounding tasks, we use the mAP metric to evaluate how well the predicted bounding boxes match the ground truth. For Visual Question Answering (VQA) tasks conducted on the NuScenes-QA (Qian et al. 2024) dataset, we differentiate between the types of questions to select appropriate evaluation metrics. Questions pertaining to object categories are assessed using the Accuracy metric, which measures the proportion of correctly identified categories. In contrast, questions related to spatial attributes such as distance and displacement are evaluated using the Mean Absolute Error (MAE), which quantifies the average magnitude of errors in distance or displacement predictions.

For open-loop driving, we follow standard practices by utilizing the implementation provided by VAD (Jiang et al. 2023) to evaluate planning within 1, 2, and 3-seconds time horizons. We use two widely accepted metrics to assess performance: the L2 error, calculated by comparing the predicted trajectory of the self-vehicle with the ground truth trajectory at corresponding way-points, and the collision rate, determined by checking for any intersections between the self-vehicle and other entities in the scene.

Main Results

We evaluated our method on NuScenes-MQA (Inoue et al. 2024) in Table 4 and OmniDrive-NuScenes (Wang et al. 2024a) in Table 5, and observed significant improvements across various metrics. Specifically, in the NuScenes-MQA

dataset, ACC measures the average accuracy of yes/no questions, classification tasks, and counting tasks under correct category conditions. Our approach achieves a 10.6% improvement over the previous state-of-the-art (SoTA) methods. In the OmniDrive-NuScenes dataset, we evaluated caption-based metrics, demonstrating a 51.4% improvement in CIDEr.

As shown in Table 3, for the object-level risk assessment dataset, our evaluation is divided into three components: the language score evaluate the quality of risk explanation. Accuracy measures the precision of risk information (categories, levels, objects, and presence), where categories, levels, and objects are evaluated only when the exists is correctly. mAP evaluates the localization of the maximum risk targets. We evaluated both the baseline model and our model with different added modules, and our complete model achieved optimal performance.

Furthermore, we utilized the NuInstruct (Ding et al. 2024) dataset to assess our method's capability to handle multi-view information, where we also observed notable improvements across all metrics. The results are shown in Table 6. Specifically, we compute the average MAE for distance, speed, count and motion. We calculate the average ACC for loset object, status and is in the same road. For the risk tasks, we employ the mAP metric, and for the reasoning tasks, we use BLUE4. Our approach achieves SoTA performance in MAE, ACC, and BLUE4, with a 20.23% improvement in the ACC metric and 98.44% improvement in the BLUE4 metric. We also achieve comparable performance in the mAP metric.

For the VQA task on NuScenes-QA shown in Table 7, we also have achieved comparable performance. These experimental results robustly demonstrate the effectiveness of our proposed method.

Ablation Studies

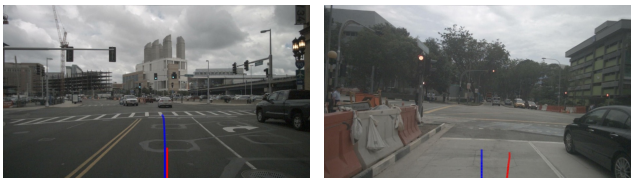
As shown in Table 8, we conducted ablation studies on the NuScenes-QA (Qian et al. 2024) and OmniDrive-NuScenes (Wang et al. 2024a) datasets to validate the effectiveness of our proposed module. And in Table 3, we present the impact of different modules on reasoning abilities under perception-limited conditions.

Experimental results demonstrate the critical role of our training strategy and dataset in enhancing the grounding task performance. When we incorporated BEV representations, there was a noticeable improvement in the grounding task's performance. However, this addition had a relatively minor impact on captioning tasks, indicating that BEV benefits are more pronounced in grounding than in captioning. Moreover, integrating the interactor component without the top-k operation yielded substantial improvements across various evaluation metrics. This enhancement is attributed to the effective integration of instruction-guided information, which considerably boosts performance. The top-k operation, which aggregates information more efficiently, further optimizes the system's capabilities. Its inclusion facilitates a more nuanced understanding by the Large Language Model (LLM), leading to the best overall performance of our complete model.

Model	ACC ↑	CIDEr ↑	BLUE4 ↑	mAP ↓
Base	64.2	70.2	36.4	0
+ TS	71.9	92.1	36.8	10.41
+ BEV	73.3	93.6	36.7	11.66
+ Interactor	74.1	101.2	41.69	13.01
+ Selection	**74.4**	**103.9**	**49.08**	**16.66**

Table 8: Ablation study on OmniDrive-NuScenes, NuScenes-MQA and NuInstruct datasets. TS represents our three stage training strategy.

(a) Turn right at night



(b) Go straight at the crossing

Figure 5: **Qualitative results with planning.** The red line represents the ground truth path, while the blue line indicates the path predicted by our method. These results were obtained without ego status.

Parameter Analyzation. We analyzed the impact of the value of k on model performance. As shown in Table 10, the optimal performance was achieved when $k = 90$. We attribute this to the redundancy present in the data. There is a significant amount of information in the visual tokens that is irrelevant to the instructions. We utilized a selection module to extract the k most relevant visual tokens. However, a value of k that is too small leads to the loss of key information, while a value that is too large fails to mitigate redundancy.

Open-loop Planning

We compare our method with previous SoTA approaches in Table 9. We adopt a distinct encoding scheme for ego status: firstly, we convert all units from meters to centimeters and round to the nearest integer to facilitate tokenization by the language model. Subsequently, ego status is input to the large language model in a linguistic format (*e.g.* "Given the ego status: lateral velocity is 0 cm/s; longitudinal velocity is 418 cm/s; lateral acceleration is 5 cm/s^2; longitudinal acceleration is 93 cm/s^2; The ego car will TURN LEFT. Output planning results.").

As described in BEV-Planner (Li et al. 2024), encoding ego status can significantly enhance the performance of all methods. Therefore, we conducted experiments focusing on both ego status and high-level commands. Significantly, for our approach, encoding ego status substantially

METHOD	HIGH LEVEL COMMEND	EGO STATUS BEV	EGO STATUS Planner	L2 (m) ↓ 1s	2s	3s	AVG	COLLISION (%) ↓ 1s	2s	3s	AVG
UniAD	✓	✗	✗	0.59	1.01	1.48	1.03	0.16	0.51	1.64	0.77
	✓	✓	✓	0.20	0.42	0.75	0.46	0.02	0.25	0.84	0.37
VAD-Base	✓	✗	✗	0.69	1.22	1.83	1.25	0.06	0.68	2.52	1.09
	✓	✓	✓	0.17	0.34	0.60	0.37	0.04	0.27	0.67	0.33
BEV-Planner	✓	✗	✗	0.30	0.52	0.83	0.55	0.10	0.37	1.30	0.59
	✓	✓	✓	0.16	0.32	0.57	0.35	0.00	0.29	0.73	0.34
OmniDrive	✗	✗	✗	1.15	1.96	2.84	1.98	0.80	3.12	7.46	3.79
	✓	✗	✗	0.40	0.80	1.32	0.84	0.04	0.46	2.32	0.94
	✓	✓	✓	0.14	**0.29**	0.55	0.33	**0.00**	**0.13**	0.78	0.30
Ours	✗	✗	✗	0.3	0.65	1.14	0.7	0.14	0.49	1.03	0.55
	✓	✗	✗	0.29	0.6	1.03	0.64	0.1	0.2	0.51	0.27
	✓	✓	✓	**0.14**	0.3	**0.55**	**0.33**	0.07	0.14	**0.32**	**0.18**

Table 9: Comparsions on the open-loos planning. For a fair comparison, we refer to the reproduced results in BEV-Planner. The bold numbers represent the highest accuracy. The optimal results are highlighted in bold.

k	BLUE4 ↑	CIDEr ↑	ROUGE_L ↑	ACC ↑	mAP ↑
30	14.2	**105.3**	38.5	57.3	13.01
60	13.1	98.9	38.3	54.8	11.6
90	**16.9**	103.9	38.5	**57.4**	**16.66**
120	15.3	102.2	**38.6**	56.8	12.5

Table 10: Parameter analyze on OmniDrive-NuScenes, NuScenes-QA and NuInstruct datasets

improve planning performance, whereas high-level commands offer limited improvements in planning performance. Upon analysis, we consider that in the nuScenes (Caesar et al. 2019) scenario, the driving behavior (high-level commends) choices available in most scenarios are unique, and our model is capable of fully perceiving the current scene to make current plans.

Our method achieves a new SoTA performance in collision rate and reaches comparable in L2 error. Additionally, our method attains SoTA performance even without incorporating high-level commands and ego status. When employing high-level commands but omitting ego status, our method also achieves SoTA performance in collision rate and demonstrates comparable results in L2 error.

Qualitative Results

We visualized the planning results of open-loop driving without ego status to better understand the effectiveness of our approach. As shown in Figure 5, our method, while producing higher L2 errors after the training phase, demonstrates notable improvements in the quality of the planning paths generated. For example, as shown in Figure 5 (a), a larger steering angle enables the ego vehicle to complete turns more quickly. In Figure 5 (b), in the scenario with a traffic light at an intersection, the model decelerates and stops when the light is red. However, when the light is green, the model accelerates through the intersection, which differs from the ground truth.

The qualitative results reveals that our approach consistently generates paths that are more reasonable and practical compared to previous methods. This enhanced path generation capability is not merely a theoretical improvement but translates into significant practical benefits. Overall, while the L2 error did not show a significant decrease, the qualitative improvements in path planning and the substantial reduction in collision occurrences underscore the effectiveness and practicality of our method in open-loop driving scenarios.

Conclusion

In this paper, we propose a framework to integrate world-knowledge and perception ability. By combining a instruction-guided interaction module, our approach effectively fuses multi-view video data with natural language instructions, leading to enriched visual representations. Then, we collected and refined a large-scale multi-modal dataset that includes 2 million natural language QA pairs, 1.7 million grounding data. The risk assessment data validates the performance of our approach under perception-limited conditions. Extensive experiments across tasks such as VQA, open-loop driving, and detection demonstrate the effectiveness, comprehensiveness, and generalization of our approach.

Acknowledgements

This work was supported by the Natural Science Foundation of Shenzhen under Grant No. JCYJ20230807142703006, Natural Science Foundation of China (NSFC) under Grants No. 62176021 and No. 62172041, and Key Research Platforms and Projects of the Guangdong Provincial Department of Education under Grant No.2023ZDZX1034.

References

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

Alayrac, J.-B.; Donahue, J.; Luc, P.; Miech, A.; Barr, I.; Hasson, Y.; Lenc, K.; Mensch, A.; Millican, K.; Reynolds, M.; et al. 2022. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35: 23716–23736.

Awadalla, A.; Gao, I.; Gardner, J.; Hessel, J.; Hanafy, Y.; Zhu, W.; Marathe, K.; Bitton, Y.; Gadre, S.; Sagawa, S.; et al. 2023. Openflamingo: An open-source framework for training large autoregressive vision-language models. *arXiv preprint arXiv:2308.01390*.

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

Bai, Y.; Wu, D.; Liu, Y.; Jia, F.; Mao, W.; Zhang, Z.; Zhao, Y.; Shen, J.; Wei, X.; Wang, T.; et al. 2024. Is a 3D-Tokenized LLM the Key to Reliable Autonomous Driving? *arXiv preprint arXiv:2405.18361*.

Brohan, A.; Brown, N.; Carbajal, J.; Chebotar, Y.; Chen, X.; Choromanski, K.; Ding, T.; Driess, D.; Dubey, A.; Finn, C.; et al. 2023. Rt-2: Vision-language-action models transfer web knowledge to robotic control. *arXiv preprint arXiv:2307.15818*.

Caesar, H.; Bankiti, V.; Lang, A. H.; Vora, S.; Liong, V. E.; Xu, Q.; Krishnan, A.; Pan, Y.; Baldan, G.; and Beijbom, O. 2019. nuScenes: A multimodal dataset for autonomous driving. *arXiv preprint arXiv:1903.11027*.

Chen, L.; Sinavski, O.; Hünermann, J.; Karnsund, A.; Willmott, A. J.; Birch, D.; Maund, D.; and Shotton, J. 2024a. Driving with LLMs: Fusing Object-Level Vector Modality for Explainable Autonomous Driving. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*.

Chen, X.; Wang, X.; Changpinyo, S.; Piergiovanni, A.; Padlewski, P.; Salz, D.; Goodman, S.; Grycner, A.; Mustafa, B.; Beyer, L.; et al. 2022. Pali: A jointly-scaled multilingual language-image model. *arXiv preprint arXiv:2209.06794*.

Chen, Z.; Wu, J.; Wang, W.; Su, W.; Chen, G.; Xing, S.; Zhong, M.; Zhang, Q.; Zhu, X.; Lu, L.; et al. 2024b. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 24185–24198.

Cohen, G. H. 1997. ALIGN: a program to superimpose protein coordinates, accounting for insertions and deletions. *Journal of applied crystallography*, 30(6): 1160–1161.

Cui, Y.; Huang, S.; Zhong, J.; Liu, Z.; Wang, Y.; Sun, C.; Li, B.; Wang, X.; and Khajepour, A. 2023. Drivellm: Charting the path toward full autonomous driving with large language models. *IEEE Transactions on Intelligent Vehicles*.

Dewangan, V.; Choudhary, T.; Chandhok, S.; Priyadarshan, S.; Jain, A.; Singh, A. K.; Srivastava, S.; Jatavallabhula, K. M.; and Krishna, K. M. 2023. Talk2BEV: Language-enhanced Bird's-eye View Maps for Autonomous Driving. *arXiv preprint arXiv:2310.02251*.

Ding, X.; Han, J.; Xu, H.; Liang, X.; Zhang, W.; and Li, X. 2024. Holistic Autonomous Driving Understanding by Bird's-Eye-View Injected Multi-Modal Large Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13668–13677.

Ding, X.; Han, J.; Xu, H.; Zhang, W.; and Li, X. 2023. Hilm-d: Towards high-resolution understanding in multimodal large language models for autonomous driving. *arXiv preprint arXiv:2309.05186*.

Driess, D.; Xia, F.; Sajjadi, M. S.; Lynch, C.; Chowdhery, A.; Ichter, B.; Wahid, A.; Tompson, J.; Vuong, Q.; Yu, T.; et al. 2023. Palm-e: An embodied multimodal language model. *arXiv preprint arXiv:2303.03378*.

Fang, Y.; Sun, Q.; Wang, X.; Huang, T.; Wang, X.; and Cao, Y. 2023. EVA-02: A Visual Representation for Neon Genesis. *arXiv preprint arXiv:2303.11331*.

Fu, D.; Li, X.; Wen, L.; Dou, M.; Cai, P.; Shi, B.; and Qiao, Y. 2024. Drive like a human: Rethinking autonomous driving with large language models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 910–919.

GLM, T.; Zeng, A.; Xu, B.; Wang, B.; Zhang, C.; Yin, D.; Rojas, D.; Feng, G.; Zhao, H.; Lai, H.; Yu, H.; Wang, H.; Sun, J.; Zhang, J.; Cheng, J.; Gui, J.; Tang, J.; Zhang, J.; Li, J.; Zhao, L.; Wu, L.; Zhong, L.; Liu, M.; Huang, M.; Zhang, P.; Zheng, Q.; Lu, R.; Duan, S.; Zhang, S.; Cao, S.; Yang, S.; Tam, W. L.; Zhao, W.; Liu, X.; Xia, X.; Zhang, X.; Gu, X.; Lv, X.; Liu, X.; Liu, X.; Yang, X.; Song, X.; Zhang, X.; An, Y.; Xu, Y.; Niu, Y.; Yang, Y.; Li, Y.; Bai, Y.; Dong, Y.; Qi, Z.; Wang, Z.; Yang, Z.; Du, Z.; Hou, Z.; and Wang, Z. 2024. ChatGLM: A Family of Large Language Models from GLM-130B to GLM-4 All Tools. arXiv:2406.12793.

Inoue, Y.; Yada, Y.; Tanahashi, K.; and Yamaguchi, Y. 2024. Nuscenes-mqa: Integrated evaluation of captions and qa for autonomous driving datasets using markup annotations. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 930–938.

Jiang, B.; Chen, S.; Xu, Q.; Liao, B.; Chen, J.; Zhou, H.; Zhang, Q.; Liu, W.; Huang, C.; and Wang, X. 2023. Vad: Vectorized scene representation for efficient autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 8340–8350.

Jiang, D.; He, X.; Zeng, H.; Wei, C.; Ku, M.; Liu, Q.; and Chen, W. 2024. Mantis: Interleaved multi-image instruction tuning. *arXiv preprint arXiv:2405.01483*.

Laurençon, H.; Tronchon, L.; Cord, M.; and Sanh, V. 2024. What matters when building vision-language models? *arXiv preprint arXiv:2405.02246*.

Li, J.; Li, D.; Savarese, S.; and Hoi, S. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, 19730–19742. PMLR.

Li, Z.; Yu, Z.; Lan, S.; Li, J.; Kautz, J.; Lu, T.; and Alvarez, J. M. 2024. Is ego status all you need for open-loop end-to-end autonomous driving? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14864–14873.

Lin, J.; Yin, H.; Ping, W.; Molchanov, P.; Shoeybi, M.; and Han, S. 2024. Vila: On pre-training for visual language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 26689–26699.

Liu, H.; Li, C.; Li, Y.; Li, B.; Zhang, Y.; Shen, S.; and Lee, Y. J. 2024a. LLaVA-NeXT: Improved reasoning, OCR, and world knowledge.

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

Liu, H.; Teng, Y.; Lu, T.; Wang, H.; and Wang, L. 2023a. Sparsebev: High-performance sparse 3d object detection from multi-camera videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 18580–18590.

Liu, S.; Zeng, Z.; Ren, T.; Li, F.; Zhang, H.; Yang, J.; Li, C.; Yang, J.; Su, H.; Zhu, J.; et al. 2023b. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*.

Loshchilov, I.; and Hutter, F. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Ma, Y.; Cao, Y.; Sun, J.; Pavone, M.; and Xiao, C. 2023. Dolphins: Multimodal Language Model for Driving. *arXiv prepreint arXiv:2312.00438*.

Mei, J.; Ma, Y.; Yang, X.; Wen, L.; Cai, X.; Li, X.; Fu, D.; Zhang, B.; Cai, P.; Dou, M.; et al. 2024. Continuously Learning, Adapting, and Improving: A Dual-Process Approach to Autonomous Driving. *arXiv preprint arXiv:2405.15324*.

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

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

Sima, C.; Renz, K.; Chitta, K.; Chen, L.; Zhang, H.; Xie, C.; Luo, P.; Geiger, A.; and Li, H. 2023. DriveLM: Driving with Graph Visual Question Answering. *arXiv preprint arXiv:2312.14150*.

Sun, Q.; Cui, Y.; Zhang, X.; Zhang, F.; Yu, Q.; Wang, Y.; Rao, Y.; Liu, J.; Huang, T.; and Wang, X. 2024. Generative multimodal models are in-context learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14398–14409.

Tian, R.; Li, B.; Weng, X.; Chen, Y.; Schmerling, E.; Wang, Y.; Ivanovic, B.; and Pavone, M. 2024a. Tokenize the World into Object-level Knowledge to Address Long-tail Events in Autonomous Driving. *arXiv preprint arXiv:2407.00959*.

Tian, X.; Gu, J.; Li, B.; Liu, Y.; Zhao, Z.; Wang, Y.; Zhan, K.; Jia, P.; Lang, X.; and Zhao, H. 2024b. DriveVLM: The Convergence of Autonomous Driving and Large Vision-Language Models. *arXiv preprint arXiv:2402.12289*.

Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Wang, S.; Yu, Z.; Jiang, X.; Lan, S.; Shi, M.; Chang, N.; Kautz, J.; Li, Y.; and Alvarez, J. M. 2024a. OmniDrive: A Holistic LLM-Agent Framework for Autonomous Driving with 3D Perception, Reasoning and Planning. *arXiv preprint arXiv:2405.01533*.

Wang, T.; Xie, E.; Chu, R.; Li, Z.; and Luo, P. 2024b. Drivecot: Integrating chain-of-thought reasoning with end-to-end driving. *arXiv preprint arXiv:2403.16996*.

Wang, W.; Xie, J.; Hu, C.; Zou, H.; Fan, J.; Tong, W.; Wen, Y.; Wu, S.; Deng, H.; Li, Z.; et al. 2023. Drivemlm: Aligning multi-modal large language models with behavioral planning states for autonomous driving. *arXiv preprint arXiv:2312.09245*.

Wen, L.; Fu, D.; Li, X.; Cai, X.; Ma, T.; Cai, P.; Dou, M.; Shi, B.; He, L.; and Qiao, Y. 2023. Dilu: A knowledge-driven approach to autonomous driving with large language models. *arXiv preprint arXiv:2309.16292*.

Yu, J.; Wang, X.; Tu, S.; Cao, S.; Zhang-Li, D.; Lv, X.; Peng, H.; Yao, Z.; Zhang, X.; Li, H.; et al. 2024. KoLA: Carefully Benchmarking World Knowledge of Large Language Models. In *The Twelfth International Conference on Learning Representations*.