WiseAD: Knowledge Augmented End-to-End Autonomous Driving with Vision-Language Model

Songyan Zhang[1][*], Wenhui Huang[1][*], Zihui Gao[2], Hao Chen[2], Chen Lv[1][†]

[1] Nanyang Technology University, Singapore   [2] Zhejiang University, China

Abstract

The emergence of general human knowledge and impressive logical reasoning capacity in rapidly progressed vision-language models (VLMs) have driven increasing interest in applying VLMs to high-level autonomous driving tasks, such as scene understanding and decision-making. However, an in-depth study on the relationship between knowledge proficiency—especially essential driving expertise—and closed-loop autonomous driving performance requires further exploration. In this paper, we investigate the effects of the depth and breadth of fundamental driving knowledge on closed-loop trajectory planning and introduce WiseAD, a specialized VLM tailored for end-to-end autonomous driving capable of driving reasoning, action justification, object recognition, risk analysis, driving suggestions, and trajectory planning across diverse scenarios. We employ joint training on driving knowledge and planning datasets, enabling the model to perform knowledge-aligned trajectory planning accordingly. Extensive experiments indicate that as the diversity of driving knowledge extends, critical accidents are notably reduced, contributing 11.9% and 12.4% improvements in the driving score and route completion on the Carla closed-loop evaluations, achieving state-of-the-art performance. Moreover, WiseAD also demonstrates remarkable performance in knowledge evaluations on both in-domain and out-of-domain datasets.

1. Introduction

With the advancements of related modules including perception, prediction, planning, control, *etc*, autonomous driving has made significant progress in recent years, transitioning from the traditional rule-based system [1] to the end-to-end solution[13–15]. Despite the impressive breakthroughs achieved across various benchmarks, autonomous driving still faces challenges in scene understanding and struggles to leverage fundamental driving knowledge for reliable trajectory planning as a mature human driver, which may potentially hinder further development.

Recently, the emergent general intelligence exhibited by vision-language models (VLMs) [3, 5, 7, 8, 23, 26] has demonstrated a remarkable ability to comprehend visual content and perform sophisticated vision-language dialogues based on the visual and textual inputs. This suggests a potential solution for enhancing autonomous driving to emulate human drivers more closely. However, exploiting this general intelligence and harnessing the logical reasoning capabilities for trustworthy trajectory planning is non-trivial. The primary challenge is twofold: (1) *Shortage of driving-oriented knowledge in VLMs.* Following [19], we refer to the concretization and generalization of human representation of driving scenes, driving experiences, and causal reasoning as fundamental driving knowledge. The widely used VLMs are primarily designed to develop a broad cognitive understanding of the world. It has been demonstrated [21, 28] that the direct application of vanilla VLMs to answer driving-related questions leads to redundant and meaningless correspondence. (2) *Shortage of knowledge alignment for trajectory planning.* Given a target waypoint as the guidance, the task of trajectory planning is to formulate a reasonable path to reach the destination. Although pioneering works [14, 15, 34–36] have investigated the integration of various modules such as perception and prediction, and exploited advantages of multi-modal sensor fusion, the learning of navigation focuses on imitating the driving behavior of pre-defined agents while neglecting the essential driving knowledge behind. For example, autonomous vehicles may decelerate and adopt cautious driving behaviors in areas with roadside parking vehicles. However, they still struggle with understanding that these decisions are intended to prevent the sudden appearance of pedestrians, thereby avoiding collisions and ensuring safety, highlighting the need for further exploration of explicit knowledge embedding.

In this paper, we aim to tackle these two challenges by proposing WiseAD, a vision-language model tailored for autonomous driving with extensive fundamental driving knowledge covering scene understanding, object recognition, po-

---

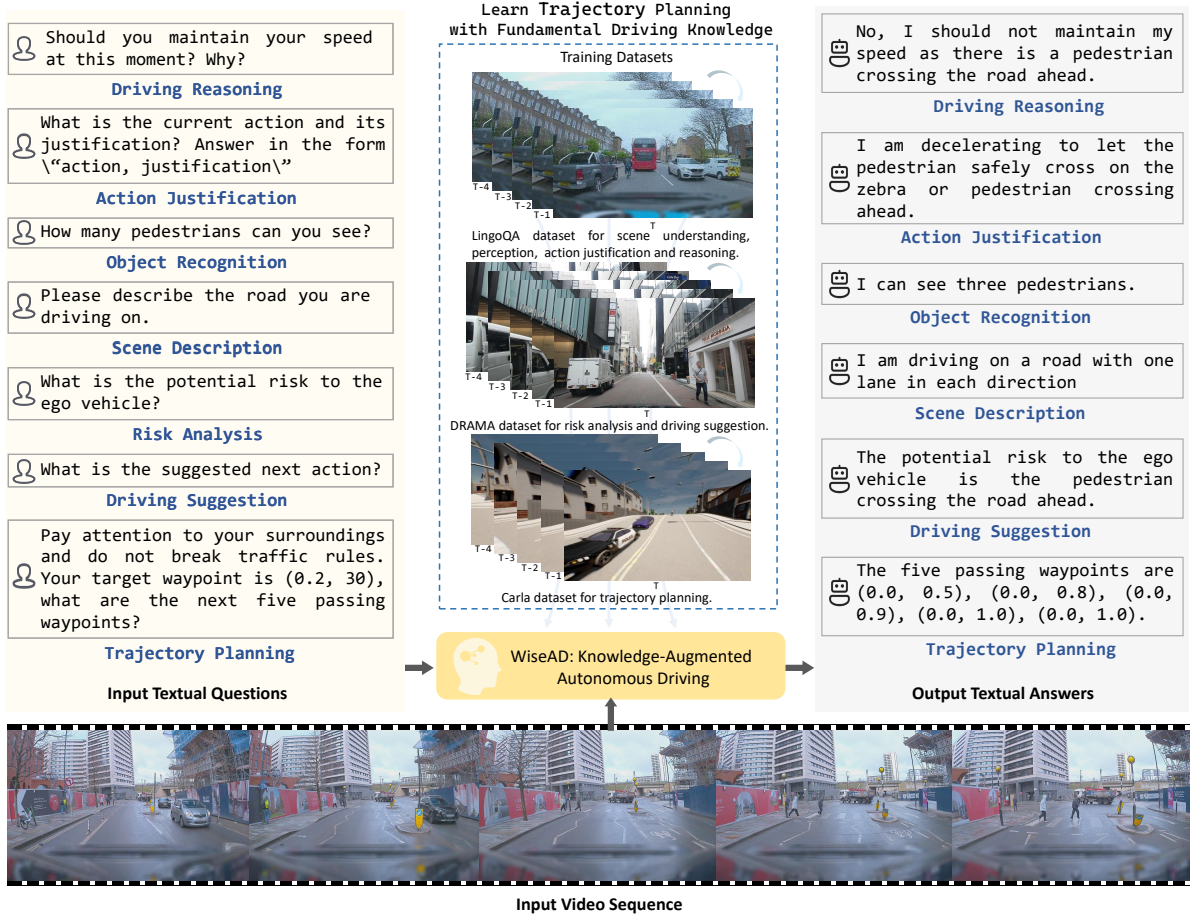[*]Co-first authors
[†]Corresponding author

Figure 1. **An overview of the proposed WiseAD**, a specialized vision-language model for end-to-end autonomous driving with extensive fundamental driving knowledge. Given a clip of the video sequence, our WiseAD is capable of answering various driving-related questions and performing knowledge-augmented trajectory planning according to the target waypoint.

tential risk analysis, driving action reasoning, driving action suggestion, and is capable of planning trajectory according to the learned knowledge. An overview of our WiseAD is illustrated in Fig 1. MobileVLM(1.7B) [8] is leveraged as our vision-language model which is a lightweight yet efficient framework targeted for mobile scale. To extend the driving-oriented essential knowledge, we first collect video question-answering datasets including LingoQA [28], and DRAMA [27], expanding both knowledge depth (diverse scenarios) and knowledge width (various tasks). To align trajectory planning with driving expertise, we integrate the knowledge and trajectory planning data for joint learning, which enables the model to learn how to infer the future trajectory and why such a path is planned. Furthermore, to seamlessly incorporate the linguistic capabilities of vision-language models, the representation of planned trajectories is also unified under textual scope as DriveVLM[39].

Comprehensive experiments are conducted to demon-

strate that our WiseAD efficiently enhances trajectory planning with essential driving knowledge, achieving a significant improvement in driving score and route completion along with significantly reduced vital accidents such as collisions and running traffic lights. Besides, as the wisdom extends, question-answering capability about driving obtains an obvious promotion for both in-domain and out-of-domain knowledge evaluation. Our contributions can be summarised as follows:

- We propose WiseAD, a knowledge-augmented vision-language model tailored for autonomous driving. This model incorporates extensive foundational driving knowledge collected from diverse scenarios, enhancing general driving-oriented cognition in areas such as driving reasoning, action justification, object recognition, scene description, risk analysis, driving recommendations, and trajectory planning.
- Through extensive experiments, we demonstrate that ex-

panding the depth and breadth of knowledge with a rationale training paradigm consistently improves both knowledge evaluation outcomes and end-to-end driving performance.

- Our comprehensive experiments demonstrate the effectiveness of our WiseAD on both closed-loop driving and driving-related knowledge evaluations, achieving state-of-the-art performance.

2. Related Works

2.1. LLMs and VLMs for Autonomous Driving

The pioneering work ADAPT [16] made the early exploration to leverage the video swin transformer [24] for textual driving narration and reasoning, which provides an explicit explanation of driving behavior. DriveGPT4[43] shares a similar idea of using VLM for interpretable end-to-end driving with extensive training data. LMDrive [36] proposes an end-to-end autonomous driving model based on LLaVA [23] to process multi-modal sensor data with natural language instructions. In DriveVLM [39], a slow-fast hybrid system for autonomous driving is proposed where VLM is responsible for scenario understanding and planning enhancement. Another traditional pipeline is also integrated to meet the real-time inference requirement. DriveMLM [42] incorporates additional lidar data and proposes a multimodal model based on LLaMA [40] to provide high-level driving decisions. Instead of providing an end-to-end solution, RAG-Driver [44] uses the VLM for knowledge retrieval and enhanced generalizable driving explanations. ELM [46] integrates multiple driving tasks like object detection, activity prediction, tracking, and scenario description. The limitation of ELM is that the proposed VLM agent couldn't provide either future trajectories or driving decisions which hinders further exploration of closed-loop driving performance.

2.2. Knowledge-Augmented Dataset for Autonomous Driving

As discussed in [20], autonomous driving is gradually evolving into knowledge-driven technologies, which is greatly attributed to the emergence of knowledge-augmented datasets. Compared with traditional driving datasets [4, 9, 12, 38] with standard annotations for perception and other tasks, knowledge-augmented datasets normally introduce textual captions for explicit expertise expression. BDD-X [17] is proposed for trustworthy and user-friendly autonomous driving. It is composed of over 77 hours of videos with additional textual justifications for driving actions and has been widely used for evaluating vehicle control, explanation generation, and scene captioning. HAD dataset [18] is collected from HDD dataset [33] and contains 5675 driving video clips to provide explicit driving advice annotated by humans. The

driving suggestion covers speed, traffic conditions, road elements, and driving maneuvers. Safety has always been a critical challenge for autonomous driving, and DRAMA dataset [27] aims to provide explicit risk analysis in terms of object and scenario level with accompanying textual driving suggestions. Recently, the proposal of CODA-LM [22] dataset collects various long-tail corner cases and provides textual annotations for general perception, regional perception, and driving suggestions. NuScenes dataset [4] is a popular dataset with abundant annotations for perception, prediction, and planning tasks and has been broadly adopted in traditional solutions. Recently, some explorations have been conducted to provide NuScenes dataset with textual knowledge annotations. DriveLM [37] proposes a graph-style structure to connect the question-answering pairs across perception, prediction, and planning tasks. Instead of using video clips, only keyframes are selected. Talk2Car [10] and NuScenesQA [31] datasets are also built on Nuscenes Dataset while the former dataset focuses on converting driving commands and the latter dataset concentrates on the manual construction of scene graphs and questions by leveraging existing 3D detection annotations. In LingoQA [28], authors explore a truthfulness classifier named Lingo-Judge with a higher correlation coefficient to human evaluations. Besides, a comprehensive video question-answering dataset is proposed including tasks of driving reasoning, object recognition, action justification, and scene description. The introduction of CoVLA dataset [2] encompasses 10,000 video clips incorporating language captions describing the driving scenarios as well as the future trajectory actions. In this work, we focus on the video-based question-answering data pairs and leverage LingoQA, DriveLM, DRAMA datasets for learning fundamental driving knowledge.

3. Methodolgy

3.1. Overview of WiseAD

Our proposed WiseAD is a specialized vision-language model with extensive fundamental driving knowledge tailored for autonomous driving, capable of scene description, object recognition, action justification, potential risk analysis, driving suggestions, and trajectory planning. The output is aligned to textual space as DriveVLM[39] so that the linguistic capability from the pre-trained model can be well preserved.

Our proposed WiseAD is built upon MobileVLM[8], a computation-friendly vision-language model targeted for mobile devices. The overall framework is illustrated in Fig.2, consisting of a frozen CLIP ViT-L/14 [32] with a learnable projector for visual tokens extraction and a large language model MobileLLaMA for textual questioning and answering. Particularly, given a video sequence of T images $\mathbf{X}_v \in \mathbb{R}^{T \times H \times W \times 3}$, the CLIP ViT features
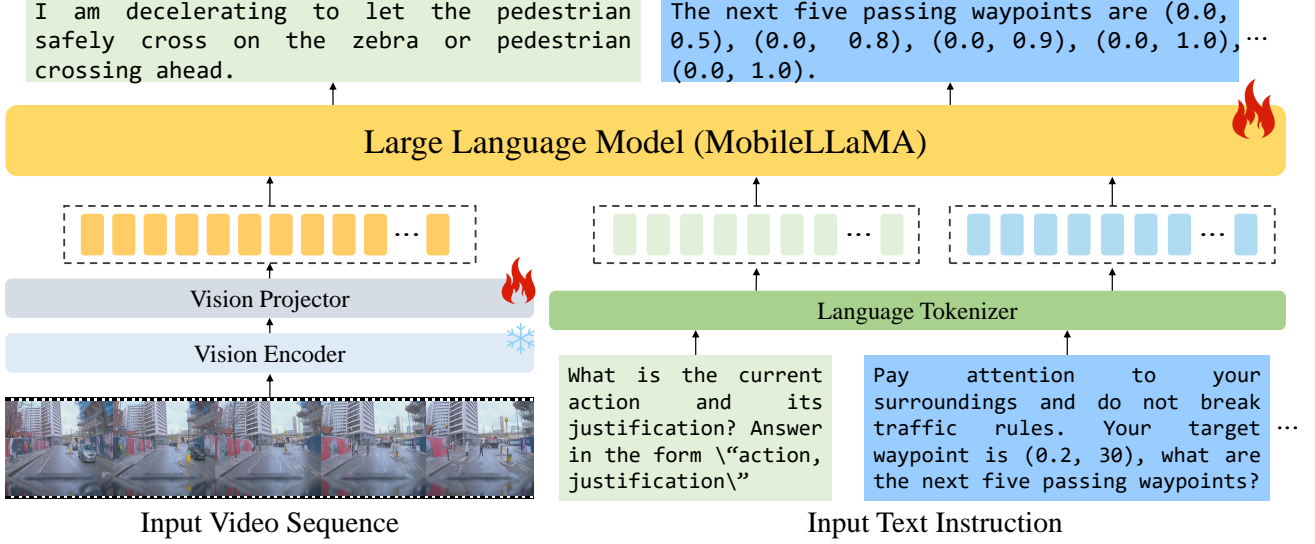
Figure 2. **The framework of the WiseAD**. Our model is built upon the MobileVLM and takes video sequences and textual prompts as input. The output for corresponding answers is unified into the linguistic expression to leverage the logical reasoning capability in vision-language models.

$\mathbf{F}_v \in \mathbb{R}^{T \times N_v \times D_v}$ are projected to modality-aligned visual tokens $\mathbf{H}_v \in \mathbb{R}^{T \times \frac{N_v}{4} \times D_l}$ where D_v and D_l denote hidden dimension of ViT and MobileLLaMA embeddings and $N_v = HW/14^2$. The projected visual tokens are then flattened along the temporal dimension. The language prompt $\mathbf{X}_l$ is tokenized to text tokens $\mathbf{H}_l \in \mathbb{R}^{N_l \times D_l}$ following the concatenation with $\mathbf{H}_v$, where N_l is the textual sequence length. The large language model takes the multimodal tokens and generates the corresponding textual response $\mathbf{Y}_a$ of length L via autoregression:

$$p(\mathbf{Y}_a|\mathbf{H}_v, \mathbf{H}_l) = \prod_{i=1}^{L} p(y_i|\mathbf{H}_v, \mathbf{H}_l, y < i), \quad (1)$$

where $p(\mathbf{Y}_a)$ is the probability of the target answers $\mathbf{Y}_a$. For closed-loop driving inference, the generated textual waypoints are converted to the numerical format. Two PID controllers are employed to adjust the steer, throttle, and braking for tracking the heading and velocity as LMDrive [36].

3.2. Data Construction

High-quality data plays a critical role in training vision-language models. In this subsection, we will discuss the formulation of training data for fundamental driving knowledge and trajectory planning.

Fundamental Driving Knowledge: A mature and trustworthy human driver makes reliable decisions based on accumulated historical information. To emulate this, we collect video-based datasets including LingoQA [28], DRAMA [27] for knowledge learning, and use BDDX [17], DriveLM [37],

and HAD [18] datasets for knowledge evaluation. For the LingoQA dataset, we follow the default configuration, where each data pair consists of 5 consecutive frames accompanied by questions and answers about driving reasoning, action justification, object recognition, and scene description. We split the original DRAMA dataset into two sets of driving suggestions and potential risk analysis to explore the effectiveness of introducing additional knowledge domains and scenarios. In DRAMA, BDDX, and HAD datasets, the original video sequences are segmented into 5 frames with evenly spaced sampling intervals. We reformulate questions using fixed question templates based on the original textual descriptions. For the DRAMA dataset, questions are constructed as "*What is the potential risk in the current scenario?*" and "*What is the suggested next action?*". For the BDDX dataset, the question is "*What is the action of the ego car?*". The corresponding answers remain unchanged as the default description in the original datasets. Fixed question template "*What the driver should pay attention?*" is used for the HAD dataset to reflect the knowledge acquisition of driving attention, which is a close task to potential risk analysis. DriveLM [37] dataset was constructed on the keyframe. We sample data pairs for the object recognition task and incorporate the previous 4 frames of the current timestep.

Textual Trajectory Planning: Following pioneering works [34, 36], we use the Carla simulator [11] to collect trajectories of an autopilot running across various scenarios at a constant frequency of approximately 10Hz. Trajectory planning for the next five waypoints is learned based on five adjacent frames from the first view, along with a destination waypoint that specifies the latitudinal and longitudinal

Dataset	Task	Question Form	Answer Form	Num
LingoQA*	Driving Reas., Object Recog., Action Just., Scene Desc.	Should you decrease your acceleration? State your reason.	Yes I should decelerate because the traffic light is red.	41.4w
DRAMA	Risk Anal., Driving Sugg.	What is the potential risk? What is the suggested next action?	There is pedestrian crossing the road. Maintain the current driving behavior.	3.5w
Carla	Trajectory Planning	Your target point is (x, y). What are the next five passing waypoints?	The next five passing waypoints are (x_0, y_0), (x_1, y_1) ...	28w
BDDX	Action Rec.	What is the action of ego car?	The car slows down.	500
DriveLM*	Object Rec.	What are objects to the front of the ego car?	There are many barriers, ...	500
HAD	Driving Attention.	What the driver should pay attention?	There are crossing cyclists in the driving lane.	500

Table 1. Detailed data construction for training and evaluating knowledge and trajectory planning. * indicates the questions are not fixed.

distance from the ego vehicle. At the training stage, the target waypoint is expressed as "*Your target waypoint is (x, y), what are the next five passing waypoints?*". The sign of x indicates the steering direction along the horizontal axis, where a positive value represents a right turn and a negative value represents a left turn. The corresponding answer is structured as "*The next five passing waypoints are (x1, y1), (x2, y2), (x3, y3), (x4, y4), (x5, y5).*".

At the inference stage, an attention prefix prompt is introduced as "*Pay attention to your surroundings and do not violate traffic rules. Your target waypoint is (x, y), what are the next five passing waypoints?*". This attention prefix serves as a trigger to leverage the learned knowledge, facilitating knowledge-augmented trajectory planning and contributing to a notable reduction in accidents. More details will be discussed in Sec 4.6.

3.3. Joint Learning for Knowledge-Augmented Trajectory Planning.

Leveraging extensive fundamental driving knowledge to enhance trajectory planning is a non-trivial challenge. An intuitive approach is to leverage the fundamental knowledge data for pre-training at the first stage, followed by the fine-tuning on the trajectory data at the second stage. However, we observed this two-stage sequential training leads to significant forgetting of driving expertise and a degeneration in navigation performance. Inspired by the human learning process where versatile intelligence and a general logical reasoning capacity contribute to the foundation of learning, we emphasize the importance of loading parameters pre-trained on large-scale data. Furthermore, during the learning process of driving, learners alternate between acquiring theoretical knowledge and applying it through practical experience. In alignment with this process, we propose to jointly learn theoretical knowledge and trajectory planning by mixing the data of these two tasks in approximately equal proportions. Given a batch of training data, the vision-language model is asked to answer driving-related questions across various

tasks while simultaneously generating a reliable route to the destination. This joint learning manner facilitates the understanding of essential knowledge underlying driving behaviors. As the diversity of fundamental knowledge accumulates, there is a notable reduction in critical accidents like vehicle collisions, coupled with an increased ratio of route completion. More details about experimental results will be discussed in Sec 4.

4. Experiment

4.1. Data Analysis

Our proposed WiseAD is trained on a mixture of various datasets including the LingoQA [28], DRAMA [27] for learning fundamental driving knowledge, along with learning trajectory planning on the Carla [11] dataset. The detailed number of data pairs and their corresponding functions for each dataset is illustrated in Tab.1. During an epoch of training, the Carla dataset is sampled twice for a balanced proportion of theoretical knowledge and trajectory planning.

We adopt the configuration in LMDrive [36] for sampling trajectory data with a pre-defined rule-based agent in the Carla simulator. It's worth mentioning that only first-view images are collected in consistency with the data format in knowledge datasets. During the evaluation stage, the LingoQA validation dataset, BDDX [17], DriveLM [37], and HAD [18] datasets are employed to validate different knowledge acquisition through QA tasks, while trajectory planning is evaluated through zero-shot testing in the Town05 environment of CARLA within a closed-loop sense. We randomly sample 500 data pairs for each of BDD-X, DriveLM, and HAD datasets to construct the zero-shot evaluation for action justification, object recognition, and driving attention, respectively.

4.2. Implementation Details

We leverage the training framework in MobileVLM [8] and initialize the model with parameters of instruction tuning. The whole training process on the mixture data takes 2

Datasets	Pretrained model	Carla Closed-Loop Eval						LingoQA Eval↑
		Driving score↑	Route compl↑	Infrac. score↑	Red light infraction↓	Collision vehicle↓	Agent blocked↓	
Sequential Training								
Carla	MobileVLM	62.46	83.47	0.74	2.60	2.35	0.14	13.4
Carla	LingoQA-Pre	52.30	75.11	0.71	2.78	2.85	3.87	12.8
Joint Training								
Carla+LingoQA	MobileVLM	63.80	87.96	0.72	3.79	5.60	0.64	58.2
Carla+LingoQA +DRAMA Suggestion	MobileVLM	66.02	89.50	0.75	2.26	1.87	1.37	58.4
Carla+LingoQA +DRAMA Suggestion +DRAMA Risk	MobileVLM	**69.88**	**93.79**	**0.76**	**2.14**	**1.43**	**0.14**	**60.4**

Table 2. Experiment on data diversity and training recipe. The best performance is reported in **bold**. The increasing depth and width of training data introduces consistent improvement with joint learning of both trajectory planning and driving knowledge.

Dataset	LingoQA↑			BDDX-Action↑		DriveLM-Obj↑		HAD-Attention↑	
	L-Judge	BLEU	CIDEr	BLEU	METEOR	BLEU	METEOR	BLEU	METEOR
C	13.4	2.2	21.5	0.0	2.2	4.4	4.0	0.0	0.2
C+L	58.2	13.7	64.3	0.3	3.5	16.9	14.9	0.8	5.9
C+L+D	**60.4**	**14.2**	**68.3**	**1.5**	**10.8**	**16.9**	**15.5**	**0.9**	**6.1**

Table 3. Experiment on the effectiveness of extending training data on knowledge evaluation. C, L, D are short for Carla, LingoQA, DRAMA datasets, respectively. The best performance is reported in **bold**.

epochs with the peak learning rate of 4×10^{-5} for the first epoch and 1×10^{-5} for the second epoch. The cosine strategy is adopted for both epochs accompanied by the warming-up ratio of 0.03 and 0.1, respectively. We use the AdamW [25] optimizer and a global batch size of 128 on 4 NVIDIA Tesla A100 (40GB) GPUs.

4.3. Evaluation Metric

For driving-related knowledge evaluation, we follow previous works [23, 28, 46] and report two established metrics of CIDEr [41], BLEU [29]. Moreover, we leverage the Lingo-Judge in LingoQA which is a pretrained transformer-based text classifier to evaluate the knowledge proficiency on the LingoQA dataset. Given a question, the human's answer, and the vision-language model's prediction, the Lingo-Judge estimates the probability that the model's answer is correct. For closed-loop evaluation on the Carla simulation, we consider three primary metrics including route completion (RC), infraction score (IS), and driving score (DS). The route completion depicts the percentage of the route that has been completed. The infraction score indicates infractions triggered by the agent. The driving score is a comprehensive metric calculated by weighting the route completion and

infraction scores. More details can be found in the Carla LeaderBoard [11]. Additionally, we report the number of routes where crucial accidents of running red lights, and collisions occur to demonstrate the effectiveness of fundamental knowledge regulation.

4.4. Impact of Knowledge Depth and Breadth

To investigate the effectiveness of driving knowledge depth and breadth, we first explore the training paradigm and conduct step-by-step incremental experiments. The baseline is learning trajectory planning based on the vanilla MobileVLM model, which is trained on large-scale versatile datasets and thus equipped with general intelligence. We start with the sequential training, first fine-tuning the MobileVLM with the LingoQA dataset and then continually learning trajectory planning. The training process and corresponding closed-loop driving performance are presented in Tab. 2. The first two rows indicate that general intelligence offers a reasonable foundation for scene understanding and planning. However, further fine-tuning with the LingoQA dataset results in a notable decline in VLM performance, particularly in closed-loop driving tasks, due to catastrophic forgetting—a common issue in continual learning.

Method	LingoQA↑			BDDX-Action↑		DriveLM-Obj↑		HAD-Attention↑	
	L-Judge	BLEU	CIDEr	BLEU	CIDEr	BLEU	CIDEr	BLEU	CIDEr
LLaVA1.5-7B [23]	38.0	4.0	32.3	**2.0**	7.4	7.1	8.6	0.4	0.0
MobileVLM-1.7B [8]	40.0	2.8	25.2	2.1	8.8	15.6	**16.4**	0.0	0.2
InternVL2-2B [5]	42.6	2.2	29.0	0.9	6.4	1.9	0.0	0.2	0.0
InternVL2-8B [5]	53.4	3.0	33.3	0.6	5.7	2.7	0.4	0.2	0.0
DeepseekVL-7B [26]	46.4	2.9	32.5	0.2	4.3	3.5	10.5	0.3	0.0
WiseAD-1.7B(ours)	**60.4**	**19.9**	**68.3**	1.5	**10.8**	**16.9**	15.5	**0.9**	**6.1**

Table 4. Comparison with other state-of-the-art methods on driving knowledge evaluation. The best performance is reported in **bold**.

Methods	Input View	DS ↑	RC ↑
TransFuser [30]	Multiview	54.52	78.41
NEAT [6]	Multiview	58.70	77.32
Roach [45]	BEV	65.26	88.24
ST-P3[13]	Multiview	55.14	86.74
VAD[15]	Multiview	64.29	87.26
WiseAD (ours)	Firstview	**69.88**	**93.79**

Table 5. Comparisons with other SOTA methods on the Carla dataset for closed-loop evaluation where our WiseAD achieves the best performance. DS is short for driving score and RC is short for route completion. The best performance is reported in **bold**.

Method	Driving score↑	Route compl↑	Infrac. score↑
w/o attention prompt	66.89	85.35	**0.78**
w attention prompt	**69.88**	**93.79**	0.76

Table 6. Ablation studies on the effectiveness of attention-guided prompts. The best performance is reported in **bold**.

This observation motivates us to shift to a joint learning protocol. More specifically, we simultaneously train fundamental driving knowledge and trajectory planning on the pre-trained MobileVLM weights, enhancing both closed-loop driving performance and knowledge acquisition. As shown in the third row, joint training significantly improves scene understanding, as validated by LingoQA evaluation, and modestly enhances driving performance, specifically in route completion and driving score. However, we observe an increase in undesirable behaviors, such as running red lights and colliding with other vehicles, likely due to extended driving distances—highlighting a need for enhanced traffic rule compliance and safety considerations. To address this, we introduced additional domain knowledge from the DRAMA dataset, focusing on driving suggestions to mitigate these behaviors. Although the improvement in LingoQA performance is minor, traffic rule violations and hazardous behaviors decrease substantially, enhancing all three primary driving metrics. Finally, we achieved further improvements across all metrics by incorporating risk analysis data as additional knowledge on top of prior domains. Compared to training solely on the CARLA dataset, joint learning with foundational driving knowledge improves the driving score by 11.9%, contributing to a significant drop in all critical driving accidents and, therefore, increasing driving safety.

Moreover, we also explore the effectiveness of extending training data over driving knowledge acquisition. As presented in Tab.3, zero-shot evaluations on the BDDX, DriveLM Object, and HAD Attention datasets witness an increasing improvement in terms of the all the metrics. Overall, expanding knowledge depth (data amount) and breadth (diverse domains) with the rationale training recipe results in consistent enhancements in both knowledge evaluations and end-to-end driving performance.

4.5. State-of-the-Art Benchmark

To further validate the positive impact of knowledge augmentation, we compare WiseAD with several representative VLMs equipped only with general knowledge. In addition to the LingoQA test dataset, we employ three additional datasets—BDDX, DriveLM, and HAD—for zero-shot evaluation. As shown in Tab.4, WiseAD achieves the highest overall performance on the LingoQA dataset, surpassing other VLMs by a significant margin, even with a smaller model size. Furthermore, its superior performance in zero-shot evaluation demonstrates WiseAD's advantage, providing essential driving priors to execute subsequent closed-loop driving tasks. Finally, qualitative validation results on randomly selected driving scenarios are presented in Fig.3, underscoring the importance of integrating fundamental driving knowledge on top of general intelligence.

As for the closed-loop benchmark, we report the driving performance of zero-shot evaluation on Town05. As shown in Tab.5, with enhanced driving knowledge, our WiseAD achieves the best zero-shot driving performance on closed-

Object Recognition

Q:Is the motorcyclist in your lane?
Internvl-8B: No, the motorcyclist is not in your lane. They are in the opposite lane, moving away from you.
Ours: Yes, the motorcyclist is in my lane.
GT: Yes.

Risk Analysis

Q: What is the potential risk to the ego vehicle?
Ours: The potential risk to the ego vehicle is the cyclist ahead of me.
GT: There is a cyclist in front.

Drive Reasoning

Q: Do you need to come to a stop? What factors contribute to this decision?
GT: No, but I need to proceed cautiously. The pedestrian on the left might start crossing the road, and the cyclist might decide to proceed straight on my lane.
Ours: No, I don't need to stop as there are no obstacles or hazards in my way.
Internvl-8B: Yes, you need to come to a stop. The factors contributing to this decision include the presence of pedestrians crossing the road, the cyclist in the bike lane, and the green bus approaching from the opposite direction. These elements indicate that it is necessary to stop to ensure the safety of all road users.

Action Justification

Q: What is the current action and its justification? Answer in the form \"action, justification\".
Internvl-8B: The current action is the pedestrian crossing the street. The justification for this action is that the pedestrian is using the designated pedestrian crossing, which is a safe and legal way to cross the road. The pedestrian is also following traffic rules by looking both ways and waiting for a safe moment to cross
Ours: I am decelerating to let the pedestrian safely cross on the zebra or pedestrian crossing ahead.
GT: The car stops, as a pedestrian is engaged at the zebra crossing.

Figure 3. Qualitative comparison with the InternVL-8B for driving knowledge evaluation.

loop evaluation. Compared with previous representative methods based on data-driven, the metric of route completion is improved by 21% at most. For the comprehensive metric, the driving score, WiseAD outperforms a certain margin against the pioneering end-to-end autonomous driving works VAD [15] and ST-P3 [13].

4.6. Ablation Study on Attention-Prefix Prompt

As discussed in Sec.3.2, we add an attention-based prefix to the target waypoint prompt to explicitly guide the model in leveraging fundamental driving knowledge. As shown in Tab.6, removing the prefix "Pay attention to your surroundings and do not break traffic rules" leads to a significant performance drop, with route completion decreasing from 93.79 to 85.35 and the driving score declining from 69.88 to

66.89. This result validates that WiseAD effectively understands the textual guidance and has the capability to align trajectory planning with learned driving knowledge.

5. Conclusion

In this work, we have presented WiseAD, a specialized VLM tailored for end-to-end autonomous driving capable of executing versatile tasks including interactive trajectory planning. WiseAD demonstrates that expanding knowledge depth and breadth with the reasonable training recipe will consistently enhance knowledge evaluations and end-to-end driving performance. In-domain and out-of-domain evaluation results show that closed-loop driving behaviors can be progressively enhanced by injecting previously unlearned domain knowledge. We believe this work provides a reliable

foundation for future research focused on closed-loop performance and real-world autonomous driving applications.

References

[1] Baidu Apollo team (2017), Apollo: Open Source Autonomous Driving, howpublished = https://github.com/apolloauto/apollo, note = Accessed: 2019-02-11. 1

[2] Hidehisa Arai, Keita Miwa, Kento Sasaki, Yu Yamaguchi, Kohei Watanabe, Shunsuke Aoki, and Issei Yamamoto. Covla: Comprehensive vision-language-action dataset for autonomous driving. *arXiv preprint arXiv:2408.10845*, 2024. 3

[3] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 2023. 1

[4] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 11621–11631, 2020. 3

[5] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 24185–24198, 2024. 1, 7

[6] Kashyap Chitta, Aditya Prakash, and Andreas Geiger. Neat: Neural attention fields for end-to-end autonomous driving. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 15793–15803, 2021. 7

[7] Xiangxiang Chu, Limeng Qiao, Xinyang Lin, Shuang Xu, Yang Yang, Yiming Hu, Fei Wei, Xinyu Zhang, Bo Zhang, Xiaolin Wei, et al. Mobilevlm: A fast, reproducible and strong vision language assistant for mobile devices. *arXiv preprint arXiv:2312.16886*, 2023. 1

[8] Xiangxiang Chu, Limeng Qiao, Xinyu Zhang, Shuang Xu, Fei Wei, Yang Yang, Xiaofei Sun, Yiming Hu, Xinyang Lin, Bo Zhang, et al. Mobilevlm v2: Faster and stronger baseline for vision language model. *arXiv preprint arXiv:2402.03766*, 2024. 1, 2, 3, 5, 7

[9] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 3213–3223, 2016. 3

[10] Thierry Deruyttere, Simon Vandenhende, Dusan Grujicic, Luc Van Gool, and Marie-Francine Moens. Talk2car: Taking control of your self-driving car. *arXiv preprint arXiv:1909.10838*, 2019. 3

[11] Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. Carla: An open urban driving simulator. In *Conference on robot learning*, pages 1–16. PMLR, 2017. 4, 5, 6

[12] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research*, 32(11):1231–1237, 2013. 3

[13] Shengchao Hu, Li Chen, Penghao Wu, Hongyang Li, Junchi Yan, and Dacheng Tao. St-p3: End-to-end vision-based autonomous driving via spatial-temporal feature learning. In *Proc. Eur. Conf. Comp. Vis.*, pages 533–549, 2022. 1, 7, 8

[14] Yihan Hu, Jiazhi Yang, Li Chen, Keyu Li, Chonghao Sima, Xizhou Zhu, Siqi Chai, Senyao Du, Tianwei Lin, Wenhai Wang, Lewei Lu, Xiaosong Jia, Qiang Liu, Jifeng Dai, Yu Qiao, and Hongyang Li. Planning-oriented autonomous driving. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 17853–17862, 2023. 1

[15] Bo Jiang, Shaoyu Chen, Qing Xu, Bencheng Liao, Jiajie Chen, Helong Zhou, Qian Zhang, Wenyu Liu, Chang Huang, and Xinggang Wang. Vad: Vectorized scene representation for efficient autonomous driving. In *Proc. IEEE Int. Conf. Comp. Vis.*, pages 8306–8316, 2023. 1, 7, 8

[16] Bu Jin, Xinyu Liu, Yupeng Zheng, Pengfei Li, Hao Zhao, Tong Zhang, Yuhang Zheng, Guyue Zhou, and Jingjing Liu. Adapt: Action-aware driving caption transformer. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 7554–7561. IEEE, 2023. 3

[17] Jinkyu Kim, Anna Rohrbach, Trevor Darrell, John Canny, and Zeynep Akata. Textual explanations for self-driving vehicles. In *Proc. Eur. Conf. Comp. Vis.*, pages 563–578, 2018. 3, 4, 5

[18] Jinkyu Kim, Teruhisa Misu, Yi-Ting Chen, Ashish Tawari, and John Canny. Grounding human-to-vehicle advice for self-driving vehicles. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10591–10599, 2019. 3, 4, 5

[19] Hector J Levesque. Knowledge representation and reasoning. *Annual Review of Computer Science*, 1(1):255–287, 1986. 1

[20] Xin Li, Yeqi Bai, Pinlong Cai, Licheng Wen, Daocheng Fu, Bo Zhang, Xuemeng Yang, Xinyu Cai, Tao Ma, Jianfei Guo, et al. Towards knowledge-driven autonomous driving. *arXiv preprint arXiv:2312.04316*, 2023. 3

[21] Yanze Li, Wenhua Zhang, Kai Chen, Yanxin Liu, Pengxiang Li, Ruiyuan Gao, Lanqing Hong, Meng Tian, Xinhai Zhao, Zhenguo Li, et al. Automated evaluation of large vision-language models on self-driving corner cases. *arXiv preprint arXiv:2404.10595*, 2024. 1

[22] Yanze Li, Wenhua Zhang, Kai Chen, Yanxin Liu, Pengxiang Li, Ruiyuan Gao, Lanqing Hong, Meng Tian, Xinhai Zhao, Zhenguo Li, et al. Automated evaluation of large vision-language models on self-driving corner cases. *arXiv preprint arXiv:2404.10595*, 2024. 3

[23] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024. 1, 3, 6, 7

[24] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3202–3211, 2022. 3

[25] I Loshchilov. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 6

[26] Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Yaofeng Sun, et al. Deepseek-vl: towards real-world vision-language understanding. *arXiv preprint arXiv:2403.05525*, 2024. 1, 7

[27] Srikanth Malla, Chiho Choi, Isht Dwivedi, Joon Hee Choi, and Jiachen Li. Drama: Joint risk localization and captioning in driving. In *Proc. Winter Conf. on Appl. of Comp. Vis.*, pages 1043–1052, 2023. 2, 3, 4, 5

[28] Ana-Maria Marcu, Long Chen, Jan Hünermann, Alice Karnsund, Benoit Hanotte, Prajwal Chidananda, Saurabh Nair, Vijay Badrinarayanan, Alex Kendall, Jamie Shotton, et al. Lingoqa: Video question answering for autonomous driving. *arXiv preprint arXiv:2312.14115*, 2023. 1, 2, 3, 4, 5, 6

[29] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002. 6

[30] Aditya Prakash, Kashyap Chitta, and Andreas Geiger. Multi-modal fusion transformer for end-to-end autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7077–7087, 2021. 7

[31] Tianwen Qian, Jingjing Chen, Linhai Zhuo, Yang Jiao, and Yu-Gang Jiang. Nuscenes-qa: A multi-modal visual question answering benchmark for autonomous driving scenario. In *Proc. AAAI Conf. Artificial Intell.*, pages 4542–4550, 2024. 3

[32] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *Proc. Int. Conf. Learn. Representations*, pages 8748–8763. PMLR, 2021. 3

[33] Vasili Ramanishka, Yi-Ting Chen, Teruhisa Misu, and Kate Saenko. Toward driving scene understanding: A dataset for learning driver behavior and causal reasoning. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 7699–7707, 2018. 3

[34] Hao Shao, Letian Wang, Ruobing Chen, Hongsheng Li, and Yu Liu. Safety-enhanced autonomous driving using interpretable sensor fusion transformer. In *Conference on Robot Learning*, pages 726–737. PMLR, 2023. 1, 4

[35] Hao Shao, Letian Wang, Ruobing Chen, Steven L Waslander, Hongsheng Li, and Yu Liu. Reasonnet: End-to-end driving with temporal and global reasoning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13723–13733, 2023.

[36] Hao Shao, Yuxuan Hu, Letian Wang, Guanglu Song, Steven L Waslander, Yu Liu, and Hongsheng Li. Lmdrive: Closed-loop end-to-end driving with large language models. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 15120–15130, 2024. 1, 3, 4, 5

[37] Chonghao Sima, Katrin Renz, Kashyap Chitta, Li Chen, Hanxue Zhang, Chengen Xie, Ping Luo, Andreas Geiger, and Hongyang Li. Drivelm: Driving with graph visual question answering. *arXiv preprint arXiv:2312.14150*, 2023. 3, 4, 5

[38] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, et al. Scalability in perception for autonomous driving: Waymo open dataset. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 2446–2454, 2020. 3

[39] Xiaoyu Tian, Junru Gu, Bailin Li, Yicheng Liu, Chenxu Hu, Yang Wang, Kun Zhan, Peng Jia, Xianpeng Lang, and Hang Zhao. Drivevlm: The convergence of autonomous driving and large vision-language models. *arXiv preprint arXiv:2402.12289*, 2024. 2, 3

[40] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. 2023. 3

[41] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 4566–4575, 2015. 6

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

[43] Zhenhua Xu, Yujia Zhang, Enze Xie, Zhen Zhao, Yong Guo, Kwan-Yee K Wong, Zhenguo Li, and Hengshuang Zhao. Drivegpt4: Interpretable end-to-end autonomous driving via large language model. *IEEE Robotics and Automation Letters*, 2024. 3

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

[45] Zhejun Zhang, Alexander Liniger, Dengxin Dai, Fisher Yu, and Luc Van Gool. End-to-end urban driving by imitating a reinforcement learning coach. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 15222–15232, 2021. 7

[46] Yunsong Zhou, Linyan Huang, Qingwen Bu, Jia Zeng, Tianyu Li, Hang Qiu, Hongzi Zhu, Minyi Guo, Yu Qiao, and Hongyang Li. Embodied understanding of driving scenarios. *arXiv preprint arXiv:2403.04593*, 2024. 3, 6