CVPR
#20277

Foundation Model for Low-Altitude Coordinated Autonomous Driving

Anonymous CVPR submission

Paper ID 20277

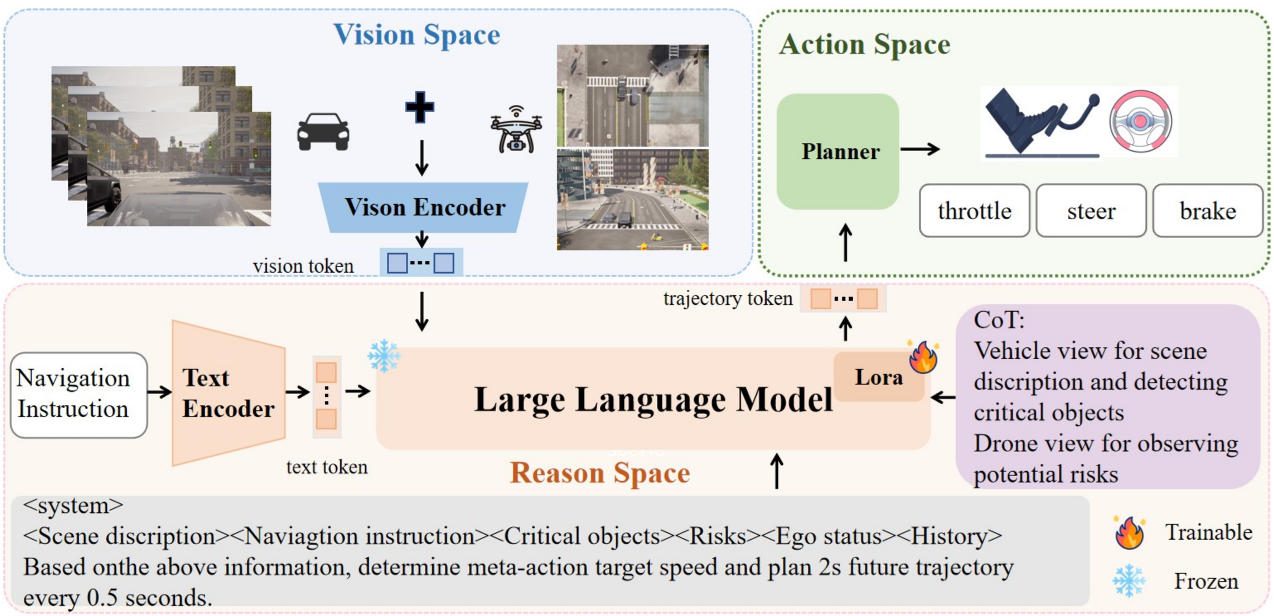

Figure 1. **The pipeline of our LALMDriver.** The low-altitude collaborative end-to-end model based on VLM aims to provide comprehensive road information, build urban agents, and improve interpretability.

Abstract

The integration of Multimodal Large Language Models (MLLMs) into autonomous driving (AD) systems represents a transformative leap in perception and reasoning. This paper introduces an end-to-end autonomous driving framework for low-altitude collaborative intelligence, leveraging Vision-Language Models (VLMs) to enhance decision-making, spatial reasoning, and trajectory planning. Our approach uniquely incorporates a drone perspective into the reasoning chain, expanding situational awareness beyond ground-level blind spots and enabling proactive risk avoidance. The drone component may function as either vehicle-mounted or standalone, providing overhead context for dynamic traffic scenes. The model generates comprehensive scene descriptions, identifies critical entities, and infers potential risks by integrating historical context and ego-state information. A closed-loop evaluation on CARLA demonstrates that our method not only improves trajectory accuracy but also significantly enhances robustness in complex and dynamic environments.

1. Introduction

Recent advances in artificial intelligence, sensor fusion, and high-performance computing have accelerated the evolution of autonomous driving (AD) technology[17, 20, 24]. Generally, AD systems can be categorized into two primary approaches: 1) Modular systerms[16, 42]that decompose into several sub-modules, such as perception[19, 27, 41], prediction[5, 13, 31], and planning[15, 29], and fixed interfaces are designed to integrate them together[4, 22]; and 2) End-to-End (E2E) autonomous driving that directly converts sensor data into control signals via a neural network, bypassing the need for symbolic interfaces and enabling

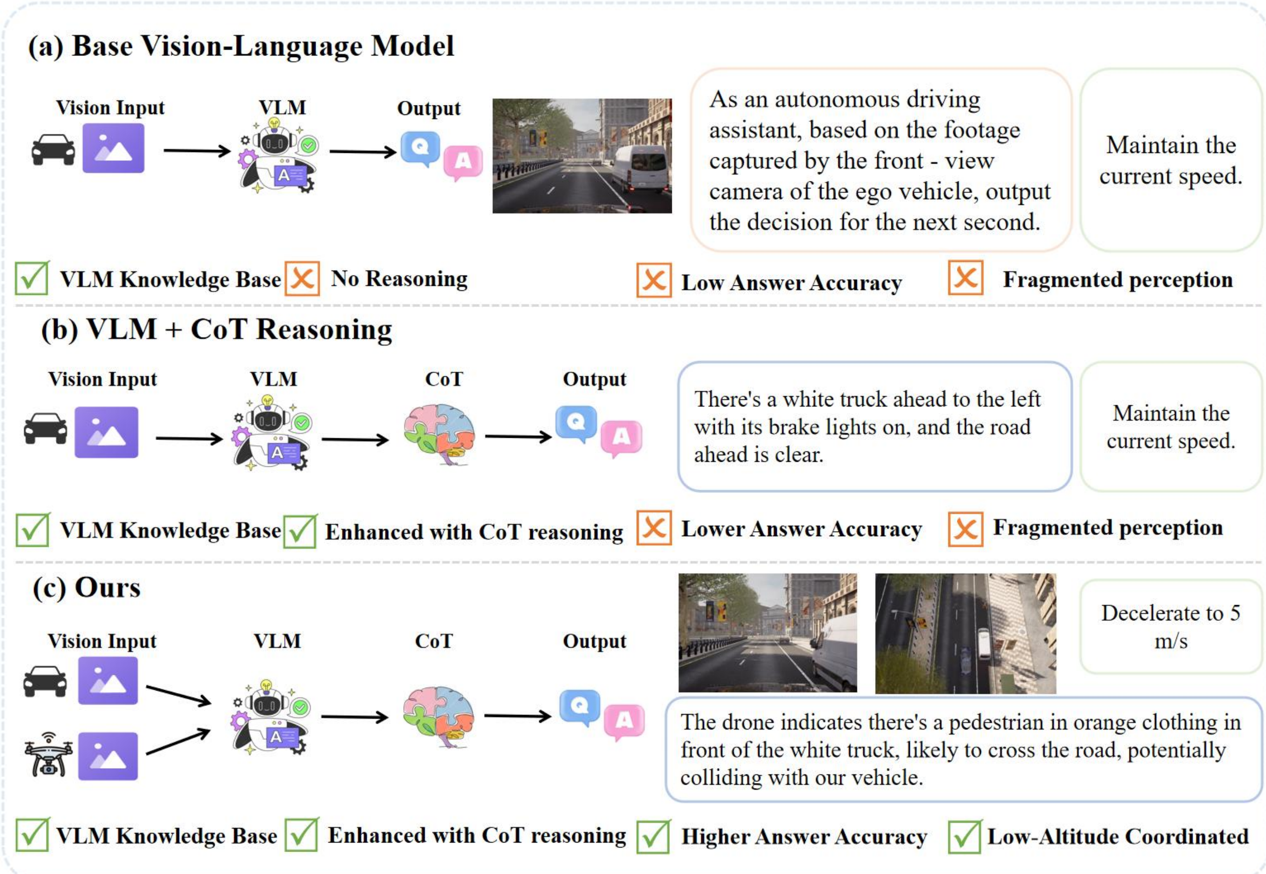

Figure 2. **Illustration of the motivation and key highlights of our proposed framework.** (a) Base VLMs use static input-output mapping with no reasoning, leading to low accuracy, frequent hallucinations and fragmented perception. (b) VLM + CoT introduces structured reasoning, improving interpretability, but still suffers from inconsistencies and lack of perception. (c) Foundation model for Low-Altitude Corrdinated Autonomous Driving (Ours) combines instructions with visual inputs from vehicle front views and UAV perspectives to improve reasoning consistency and safety.

holistic optimization.

While these approaches have achieved promising performance in open-loop evaluations[3] by imitating expert demonstrations, they often suffer from narrow generalization and limited causal reasoning.These systerms lack the ability to explain their actions and struggle in closed-loop evaluations[18] that require adaptive, interaction decision-making under uncertainty.

In contrast, Large Language Models (LLMs)[35] and Vision Language Models (VLMs) [1, 2, 8, 25]have demonstrated emergent reasoning and word-modeling capabilities approaching Artificial General Intelligence(AGI). Their ability to perform chain-of-thought (CoT) reasoning [10, 26, 37]and integrate multimodal context provides a foundation for explainable and adaptive decision-making-critical for safety and interpretability in autonomous systems.

Meanwhile, for vision-only solutions, the limitations of camera perspectives result in insufficient capability for vehicles to handle blind spots and long-range hazardous scenarios.

Our Motivation. Conventional AD systems rely on limited onboard sensors, resulting in incomplete situational awareness, especially in occluded or multi-agent environments. To address this, we propose a Low-Altitude Collaborative Autonomous Driving Framework, integrating aerial perception from UAVs with VLM-based reasoning. As illustrated in Fig. 1 and Fig. 2, our framework combines ground-view and aerial-view streams, guided by high-level navigation instructions, to form a coherent understanding of the environment.

The UAV acts as an “eye in the sky,” providing early visibility of hazards (e.g., sudden pedestrians, distant traffic jams, or occluded vehicles). This fusion of multi-perspective data enhances perception reliability, planning safety, and reasoning interpretability.

To summarize,this paper makes the following contributions:

- A low-altitude cooperative E2E framework for autonomous driving that leverages VLMs' world knowledge and reasoning capability to improve interpretability, safety, and generalization beyond traditional rule-based AD systems.
- A novel VQA-CoT dataset incorporating UAV perspectives into the reasoning process, enhancing the vehicle's understanding of occluded and dynamic regions.
- LoRA-based fine-tuning of VLM backbones (LLaVA-v1.6) for structured reasoning and trajectory generation, with superior closed-loop performance over LMDrive in complex CARLA scenarios.

2. Related Work

2.1. End-to-End Autonomous Driving

E2E[38, 43] methods process raw sensor data to output motion trajectories or low-level control signals, minimizing cumulative errors through global optimization. For instance, UniAD [11]integrates perception, prediction, and planning into a unified framework, employing query-based transformers to connect multiple tasks (detection, tracking, mapping, trajectory prediction, etc.) while optimizing computational resources. GenAD[44]and DiffusionDrive [23] explore generative models for trajectory prediction. However, these methods excel primarily in open-loop evaluation[3]. A core motivation for E2E autonomous driving is to holistically assess perception and planning as means to achieve driving objectives, rather than overfitting to perception metrics. Unlike perception, planning is inherently open-ended and challenging to quantify, necessitating closed-loop evaluation. Thus, we evaluate driving performance in CARLA[12].

2.2. LLM/VLM for E2E Driving

LLMs and VLMs demonstrate exceptional contextual reasoning and world knowledge, making them promising for autonomous driving. For example, DriveGPT4 [39] leverages GPT-4V for recognition and reasoning but struggles with numerical control signals. However, most studies rely on open-loop evaluation using simplistic datasets like nuScenes[3]. Although DriveMLM[36] and LMDrive [30] attempt closed-loop evaluation, they underperform in complex scenarios constrained by benchmarks like CARLA Town05Long.

2.3. Chain of Thought(CoT) for Visual Question Answering(VQA)

VQA-based reasoning has become a promising approach for interpretable scene understanding. In AD contexts, CoT reasoning is employed to simulate human cognitive processes. This involves recognizing key entities, predicting their future actions, and ultimately making hierarchical driving decisions, as demonstrated in DriveVLM [34]. However, conventional VQA models fail to represent multi-agent interactions or spatial hierarchies, leading to poor reasoning consistency. Graph-based extensions like DriveLM[32] improve structure but remain limited to ground-view understanding.

2.4. Collaborative Driving and UAV Integration

Vehicle-to-Everything (V2X) frameworks have improved environmental perception through infrastructure cooperation such as V2X-VLM[40]. Yet, UAV-based low-altitude collaboration remains underexplored. Unlike static roadside sensors, drones provide flexible, dynamic perspectives with broader spatial coverage. Inspired by this, we propose leveraging UAV-assisted perception fused with VLM reasoning to construct urban-scale cooperative intelligence, improving situational awareness and safety in dynamic environments.

3. Proposed Approch

Our proposed autonomous driving framework figure1 synergistically integrates multimodal VLM reasoning with aerial-ground perception fusion for closed-loop end-to-end driving.

3.1. Framework Overview

Given the time step t, we collect front-view images of the ego vehicle $I_{\text{cam,vehicle}}$, multiview UAV images $I_{\text{cam,uav}}$, the ego vehicle state S_{ego}, and the navigation information L_{nav}. The VLM outputs the trajectory and generates a clear reasoning process:

$$T_{\text{traj}}, T_{\text{reason}} = \text{VLM}(I_{\text{cam,vehicle}}, I_{\text{cam,uav}}, L_{\text{nav}}, S_{\text{ego}}) \quad (1)$$

where $T_{\text{traj}} \in \mathbb{R}^{T\times 3}$ represents the sequence of future path points and headings.

Model Input. The input consists of multimodal data, encompassing both images and text.

- Vehicle view: multi-frame front camera sequence $\{I^1_{t,\text{vehicle}}, I^2_{t,\text{vehicle}}, \dots, I^n_{t,\text{vehicle}}\}$ captured at 2 Hz
- UAV view: temporal multi-angle aerial frames $\{I^1_{t,\text{uav}}, I^2_{t,\text{uav}}, \dots, I^N_{t,\text{uav}}\}$ captured at 2 Hz
- Navigation instruction: high-level route command $\mathcal{L}_{\text{nav}}$ (e.g., "turn left after 50 m")
- Ego state: velocity and acceleration $\mathcal{S}_{\text{ego}}$

These multimodal inputs are encoded through visual and linguistic tokenizers, then fused within the VLM backbone for reasoning and trajectory generation.

Base VLM Backbone. We adopt LLaVA-v1.6-7B as vision-language backbones, due to its strong cross-modal comprehension and open-source adaptability.

CVPR
#20277

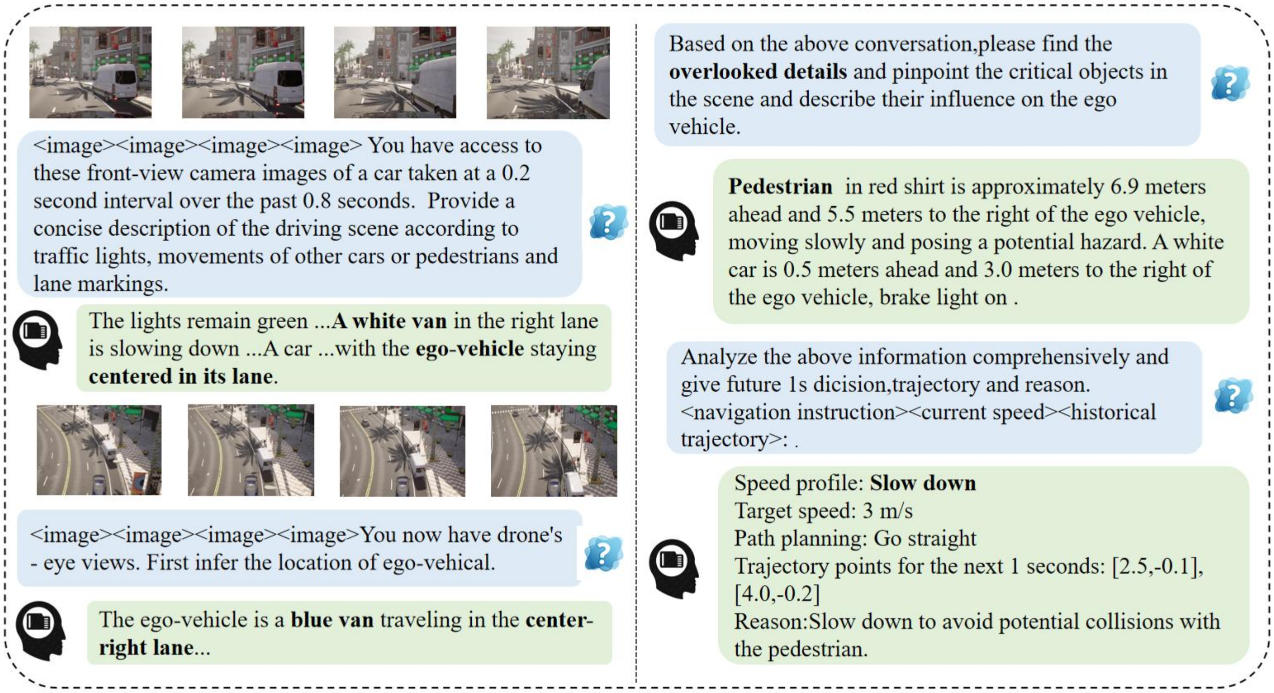

Figure 3. **Presentation of the chain of thought when building data**

Navigation instructions are first encoded in linguistic tokens $x_q \in \mathbb{R}^{L \times C}$ by a text tokenizer, where L is the length of the token and C is the dimension of the LLM. Then, the scene tokens x_s and the historical tokens x_h are combined with x_q and fed into the LLM. Moreover, we design a planning Q&A template for LLMs with a special planning token s. This accumulates the understanding and reasoning context of the entire driving scene in s, formulated as:

$$s \sim p(s|x_s, x_h, x_q, x_a), \tag{2}$$

where x_a represents the model's autoegressive answer sequence. The embedding of the planning token s serves as a condition for controlling trajectory generation.

Model Output and Control. The model outputs hierarchical planning.

- Meta-actions(A): short-term strategic primitives (e.g., accelerate, brake, change lane).
- Trajectory waypoints($W = \{w_1, w_2, \dots, w_k\}, w_i = (x_i, y_i)$):continuous motion targets sampled every Δt.
- Reason(R): reason for the decision.

Following the LBC method, to obtain the final control signals (including braking, throttle, and steering), we use two PID controllers for lateral and longitudinal control respectively to track the predicted waypoints. Lateral control adjusts the vehicle's steering, while longitudinal control regulates the vehicle's speed. PID controllers adjust based on the deviation between the desired value (predicted waypoint) and the actual value (vehicle steering and speed). The parameters used in the experiment are:

$$K_{P\_turn} = 1.25, \quad K_{I\_turn} = 0.75, \quad K_{D\_turn} = 0.3$$
$$K_{P\_speed} = 5.0, \quad K_{I\_speed} = 0.5, \quad K_{D\_speed} = 1.0$$

This approach achieves precise vehicle control through PID controllers, enabling autonomous driving.

3.2. Inference Data

Inference data provides high-quality CoT annotations, which are crucial for training VLMs with reasoning capabilities. In driving tasks, reasoning involves understanding complex semantics and interactions in dynamic environments. Despite its importance, developing high-quality, large-scale driving reasoning datasets remains a key challenge due to three main limitations: 1) limited scene diversity and repetitive examples, 2) insufficient representation of key perceptual cues (such as traffic signs and vehicle indicators), and 3) low-quality reasoning processes, such as repeatedly stopping at a stop sign without justification.

To address these issues, we propose an automatic reasoning annotation pipeline using the Kimi-1.5-671B model[33]. This pipeline can automatically generate high-criticality reasoning annotations and supports knowledge distillation from large models to more compact target

CVPR #20277

models. The pipeline generates structured reasoning annotations in four key components: detailed scene descriptions(weather, road, lane conditions), object recognition(vehicles, pedestrians, signals), prediction of intentions of surrounding key objects, and appropriate driving actions. This structured prompting approach significantly reduces nonsensical outputs and minimizes the need for manual correction.

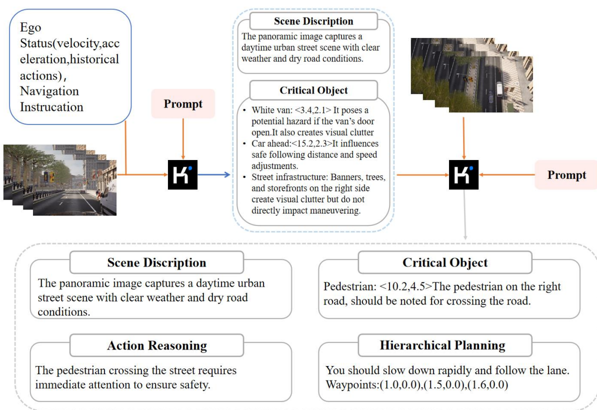

Figure 4. **The automated annotation pipeline for the data.**

3.3. Training Strategy

The training consists of two stages:1)using CoT with LoRA fine-tuning 2)training the planner

Stage1: CoT Fine-tuning with LoRA . Using LLaMA-Factory, the model is fine-tuned on CoT datasets to enhance its reasoning ability and interpretability. Each action token corresponds to 0.5 seconds of motion, and the planning horizon is set to 2 seconds. For SFT, we use a learning rate of 1×10^{-5} and the FSDP training strategy. The model is trained for 5 epochs using 1 NVIDIA 4090 GPU. We use a batch size of 1 per GPU and accumulate gradients over 4 steps. The weight parameters in the SFT loss function are set to $\lambda_a = 1$ and $\lambda_{\text{cot}} = 40$. For RFT, we adopt LoRA for parameter-efficient training. The learning rate for RFT is set to 3×10^{-5}, and the KL regularization weight β is set to 0.04. The model is fine-tuned for 6,000 steps.

Stage2:training the planner.A sequence of frames is used as input, and during training, a fixed sequence length $T_{\max}$ is set to construct batch data.During the stage, only the planner are trainable, while other components such as the vision encoder and the LLM remain frozen.When fine-tuning the modules, two loss terms are considered: an L_1 waypoint loss $\mathcal{L}_{\text{wp}}$ to ensure accurate and sensitive waypoint prediction, and a classification loss $\mathcal{L}_{\text{cls}}$ (cross-entropy) to determine whether the current frame has completed the given instruction. The overall training objective is formulated as:

$$\mathcal{L} = \mathcal{L}_{\text{wp}} + \mathcal{L}_{\text{cls}}. \tag{3}$$

A cosine learning rate scheduler is employed with an initial learning rate of 1×10^{-4}, a batch size of 32, and a total of 15 training epochs. The first 2000 iterations are used for warm-up. A weight decay of 0.07 is applied, and the maximum historical window $T_{\max}$ is set to 40. If a data segment contains more than 40 frames, it is truncated to retain only the most recent 40 frames.

4. Experiments

4.1. Experimental Setup

We evaluate our framework through closed-loop simulations in CARLA, focusing on bath custom low-altitude cooperation scenarios and the Town05-Long benchmark.This setup allows quantitative assenment of UAV-assisted reasoning and generalization crossing urban environment.

Custom Scenarios To demonstrate the UAV's contribution, we design four high-complexity scenarios (see Fig. 5):

- Ghost Pedestrian: A pedestrian suddenly burst out of a gap in parked cars on the side of the road.
- Overtaking with Occlusion: The ego vehicle overtakes safely.
- Intersection with Hidden Vehicles and VRUs: Focused on reasoning through partial visibility and vulnerable road users in intersections.
- Distant Traffic Jam or Accident: UAV height is increased to provide early detection of distant road congestion or incidents.

Benchmark Testing. Town-05 Long Benchmark is a key part of the CARLA Leaderboard. It focuses on long distance self-driving tasks on the Town-05 map. There are 10 long routes, each 1000-2000 meters with 10 intersections. It evaluates the system's ability in complex cities, covering dynamic scenarios, route accuracy, and traffic-rule compliance.

Evaluation Metrics. For custom scenarios, we briefly choose failure rate as metrics. Lower failure rate indicates higher robustness under adverse conditions. Meanwhile,we consider three key metrics established by the CARLA LeaderBoard, namely route completion (RC), infraction score (IS), and driving score (DS). RC reflects the percentage of the total route length completed along the predetermined path. Should the agent deviate excessively from the route, the trial is deemed a failure. IS quantifies infractions, including collisions and traffic violations, with the score undergoing decay via a discount factor when such occurrences happen. DS, the product of RC and IS, encapsulates both driving advancement and safety and is recognized as the principal ranking metric.

Table 1. Performance on Town05 Long benchmark on CARLA. C/L refers to camera/LiDAR. RC:route completion, IS:infraction score, DS:driving score

Method	Inference	Modality	RC↑	IS↑	DS↑
LBC [6]	CoRL20	C	31.9±2.2	0.66±0.02	12.3±2.0
Transfuser [28]	CVPR21	C&L	47.5±5.3	0.77±0.04	31.0±3.6
Roach [9]	ICCV21	C	96.4±2.1	0.43±0.03	41.6±1.8
LAV [7]	CVPR22	C&L	69.8±2.3	0.73±0.02	46.5±2.3
TCP [14]	NeurIPS22	C	80.4±1.5	0.73±0.02	57.2±1.5
LMdrive	CVPR24	C&L	78.2±3.9	0.74±0.05	57.2±2.0
LALMDriver(Ours)	-	C	85.1±4.2	0.68±0.04	58.3±1.0
ThinkTwice [21]	CVPR23	C	95.5±2.0	0.69±0.05	65.0±1.7

Table 2. Multi-Ability Results of E2E-AD Methods under baseset.

Method	Reference	Failure Rate(%)↓			
		PedestrianCrossingParkedcar	Overtaking	CrossingIntersection	Detour
LMDrive	CVPR2024	0.12	0.15	0.10	0.90
LALMDriver(Ours)	-	**0.00**	**0.09**	**0.03**	**0.06**

Table 3. **Performance comparison of 2 LLM backbones and three methods** The first method involves directly outputting control signals from the large model without using the CoT approach. The second method uses the CoT but does not incorporate the perspective of a multi-rotor drone. The third method integrates the drone's perspective into the CoT.

Module design	RC↑	IS↑	DS↑
Baseline(LLaVA-v1.6)	85.1	0.68	58.3
w/o using BEV tokens	82.9	0.64	53.0
w/o using CoT	35.1	0.41	14.4

4.2. Main Results

Our model demonstrates consistent superiority across both the CARLA Town05-Long benchmark and customized UAV-cooperative scenarios. As summarized in Table 1, LALMDriver achieves a Driving Score (DS) of 58.3, surpassing the previous closed-loop E2E methods including LMDrive (57.2 DS). The improvements mainly stem from enhanced reasoning consistency and global situational awareness brought by the UAV-assisted CoT mechanism. Notably, our method maintains a balanced Infraction Score (IS) of 0.68 while attaining a Route Completion (RC) of 85.1%, demonstrating high reliability in long-horizon navigation.

In the multi-ability evaluation shown in Table 2, LALMDriver achieves the lowest failure rate across four complex scenarios—0.00% in Pedestrian Crossing, 0.09% in Overtaking with Occlusion, 0.03% in Intersection with Hidden Vehicles, and 0.06% in Detour after Accident. These results highlight the model's strong robustness and capability to generalize across occluded, multi-agent, and long-range reasoning conditions.

Ablation analysis in Table 3 further confirms the importance of UAV integration and CoT-guided reasoning. Removing CoT reasoning causes the driving score to drop from 58.3 to 14.4, while removing UAV-derived BEV tokens decreases DS to 53.0. This indicates that the UAV perspective not only expands the visual field but also enables proactive hazard anticipation by reasoning about unseen regions.

Overall, these results demonstrate that integrating multi-perspective perception into structured reasoning significantly improves both safety and interpretability. LALMDriver effectively bridges the gap between perception and reasoning in closed-loop driving, offering a robust foundation for scalable low-altitude cooperative autonomy.

4.3. Qualitative Results

Fig. 5 shows the model's qualitative results in a typical closed-loop evaluation scenario. It displays our model's driving action reasoning and trajectory prediction outputs. We found that incorporating UAV-captured bird's-eye -view images from different angles into the chain of thought enables the model to capture the blind spots of the vehicle's front view, detect risks, and thus reason the correct causality and make right driving decisions. Then, the model predicts the planned trajectory according to the reasoning instruction, highlighting our method's impressive interpretability.

Figure 5. **Qualitative results on the closed-loop evaluation set.**

5. Conclusion

This work presents a foundation model for low-altitude collaborative autonomous driving, uniting VLM-based reasoning with UAV-assisted perception. The proposed framework enhances scene understanding, causal reasoning, and trajectory planning, offering interpretable decision-making under dynamic urban conditions.

Through closed-loop evaluations on CARLA, the system demonstrated superior performance in terms of driving scores, route completion, and infraction score. The integration of UAV views into the reasoning chain effectively mitigates occlusion, anticipates risks, and improves overall driving stability.

However, current limitations include high computational overhead for real-time deployment. Future research will explore model compression, multi-agent reasoning, and real-world UAV-vehicle co-simulation to achieve scalable deployment.

In summary, this work takes a significant step toward human-level, interpretable, and collaborative autonomous driving, bridging ground and aerial perspectives under a unified foundation model paradigm.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, and Shyamal Anadkat. Gpt-4 technical report. *arXiv preprint arXiv: 2303.08774*, 2023. 2

[2] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 2

[3] Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Anush Krishnan Qiang Xu, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. Nuscenes: A multimodal dataset for autonomous driving. In *CVPR*, pages 11621–11631, 2020. 2, 3

[4] Sergio Casas, Cole Gulino, Simon Suo, Katie Luo, Renjie Liao, and Raquel Urtasun. Implicit latent variable model for scene-consistent motion forecasting. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIII*, pages 624–641. Springer, 2020. 1

[5] Yuning Chai, Benjamin Sapp, Mayank Bansal, and Dragomir Anguelov. Multipath: Multiple probabilistic anchor trajectory hypotheses for behavior prediction. arXiv preprint arXiv:1910.05449, 2019. 1

[6] Dian Chen and Philipp Krähenbühl. Learning by cheating. In *Conference on Robot Learning (CoRL)*, 2020. 6

[7] Dian Chen, Yuke Zhou, and Philipp Krähenbühl. Learning from all vehicles. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 6

[8] 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 *CVPR*, pages 24185–24198, 2024. 2

[9] Kashyap Chitta, Aayush Prakash, and Andreas Geiger. Roach: A robust autonomy framework for autonomous driving in unstructured environments. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 6

[10] Zheng Chu, Jingchang Chen, Qianglong Chen, Weijiang Yu, Tao He, Haotian Wang, Weihua Peng, Ming Liu, Bing Qin, and Ting Liu. A survey of chain of thought reasoning: Advances, frontiers and future. arXiv preprint arXiv:2309.15402, 2023. 2

[11] UniAD contributors. Planning-oriented autonomous driving. https://github.com/OpenDriveLab/UniAD, 2023. 3

[12] Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. CARLA: An open urban driving simulator. In *Proceedings of the 1st Annual Conference on Robot Learning*, pages 1–16, 2017. 3

[13] Junru Gu, Chenxu Hu, Tianyuan Zhang, Xuanyao Chen, Yilun Wang, Yue Wang, and Hang Zhao. Vip3d: End-to-end visual trajectory prediction via 3d agent queries. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5496–5506, 2023. 1

[14] Chengyang Hu, Zhiyuan Huang, Zeyu Chen, et al. Planning-oriented autonomous driving. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 6

[15] 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 *Proceedings of the European Conference on Computer Vision*, pages 533–549, 2022. 1

[16] Yihan Hu, Jiazhi Yang, Li Chen, Keyu Li, Chonghao Sima, Xizhou Zhu, Siqi Chai, Senyao Du, Tianwei Lin, Wenhai Wang, et al. Planning-oriented autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17853–17862, 2023. 1

[17] Jyh-Jing Hwang, Runsheng Xu, Hubert Lin, Wei-Chih Hung, Jingwei Ji, Kristy Choi, Di Huang, Tong He, Paul Covington, Benjamin Sapp, Yin Zhou, James Guo, Dragomir Anguelov, and Mingxing Tan. Emma: End-to-end multimodal model for autonomous driving. arXiv preprint arXiv:2410.23262, 2024. 1

[18] Xiaosong Jia, Zhenjie Yang, Qifeng Li, Zhiyuan Zhang, and Junchi Yan. Bench2drive: Towards multi-ability benchmarking of closed-loop end-to-end autonomous driving. In *Advances in Neural Information Processing Systems*, 2024. 2

[19] Xiaohui Jiang, Shuailin Li, Yingfei Liu, Shihao Wang, Fan Jia, Tiancai Wang, Lijin Han, and Xiangyu Zhang. Far3d: Expanding the horizon for surround-view 3d object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2561–2569, 2024. 1

[20] Jinlong Li, Baolu Li, Zhengzhong Tu, Xinyu Liu, Qing Guo, Felix Juefei-Xu, Runsheng Xu, and Hongkai Yu. Light the night: A multi-condition diffusion framework for unpaired low-light enhancement in autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15205–15215, 2024. 1

[21] Qingwen Li, Zeyu Chen, Dian Chen, and Philipp Krähenbühl. Think twice before driving: Towards scalable decoders for end-to-end autonomous driving. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 6

[22] Zhiqi Li, Wenhai Wang, Hongyang Li, Enze Xie, Chonghao Sima, Tong Lu, Yu Qiao, and Jifeng Dai. Bevformer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers. In *European Conference on Computer Vision*, pages 1–18. Springer, 2022. 1

[23] Bencheng Liao, Shaoyu Chen, Haoran Yin, Bo Jiang, Cheng Wang, Sixu Yan, Xinbang Zhang, Xiangyu Li, Ying Zhang, Qian Zhang, and Xinggang Wang. Diffusiondrive: Truncated diffusion model for end-to-end autonomous driving. *arXiv preprint arXiv:2411.15139*, 2024. 3

[24] LLVM-AD Workshop Committee. Prospective of autonomous driving - multimodal llms, world models, embodied intelligence, ai alignment, and mamba. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2025. 1

[25] Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Hao Yang, Yaofeng Sun, Chengqi Deng, Hanwei Xu, Zhenda Xie, and Chong Ruan. Deepseek-vl: Towards real-world vision-language understanding, 2024. 2

[26] Rui Pan, Shuo Xing, Shizhe Diao, Wenhe Sun, Xiang Liu, Kashun Shum, Renjie Pi, Jipeng Zhang, and Tong Zhang.

Plum: Prompt learning using metaheuristic. arXiv preprint arXiv:2311.08364, 2023. 2

[27] Jonah Philion and Sanja Fidler. Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3d. In *Proceedings of the European Conference on Computer Vision*, pages 194–210, 2020. 1

[28] Aayush Prakash, Kashyap Chitta, and Andreas Geiger. Transfuser: Imitation with transformer-based sensor fusion for autonomous driving. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 6

[29] Aditya Prakash, Kashyap Chitta, and Andreas Geiger. Multimodal 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. 1

[30] 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 *CVPR*, pages 15120–15130, 2024. 3

[31] Shaoshuai Shi, Li Jiang, Dengxin Dai, and Bernt Schiele. Motion transformer with global intention localization and local movement refinement. In *Advances in Neural Information Processing Systems*, pages 6531–6543, 2022. 1

[32] 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

[33] Kimi Team, Angang Du, Bofei Gao, Bowei Xing, Changjiu Jiang, Cheng Chen, Cheng Li, Chenjun Xiao, Chenzhuang Du, Chonghua Liao, et al. Kimi k1.5: Scaling reinforcement learning with llms. *arXiv preprint arXiv:2501.12599*, 2025. 4

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

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

[36] 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

[37] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models, 2023. 2

[38] Penghao Wu, Xiaosong Jia, Li Chen, Junchi Yan, Hongyang Li, , and Yu Qiao. Trajectory-guided control prediction for end-to-end autonomous driving: A simple yet strong baseline. *arXiv preprint arXiv:2206.08129*, 2022. 3

[39] 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

[40] Junwei You, Haotian Shi, Zhuoyu Jiang, Zilin Huang, Rui Gan, Keshu Wu, Xi Cheng, Xiaopeng Li, and Bin Ran. V2x-vlm: End-to-end v2x cooperative autonomous driving through large vision-language models. *arXiv preprint arXiv:2408.09251*, 2024. 3

[41] Diankun Zhang, Zhijie Zheng, Haoyu Niu, Xueqing Wang, and Xiaojun Liu. Fully sparse transformer 3-d detector for lidar point cloud. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–12, 2023. 1

[42] Diankun Zhang, Guoan Wang, Runwen Zhu, Jianbo Zhao, Xiwu Chen, Siyu Zhang, Jiahao Gong, Qibin Zhou, Wenyuan Zhang, Ningzi Wang, et al. Sparsead: Sparse query-centric paradigm for efficient end-to-end autonomous driving. arXiv preprint arXiv:2404.06892, 2024. 1

[43] 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 (ICCV)*, 2021. 3

[44] Wenzhao Zheng, Ruiqi Song, Xianda Guo, Chenming Zhang, and Long Chen. Genad: Generative end-to-end autonomous driving. *arXiv preprint arXiv: 2402.11502*, 2024. 3