VLR-Driver: Large Vision-Language-Reasoning Models for Embodied Autonomous Driving

Anonymous ICCV submission

Paper ID 14508

Abstract

The rise of embodied intelligence and multi-modal large language models has led to exciting advancements in the field of autonomous driving, establishing it as a prominent research focus in both academia and industry. However, when confronted with intricate and ambiguous traffic scenarios, the lack of logical reasoning and cognitive decision-making capabilities remains the primary challenge impeding the realization of embodied autonomous driving. Although Vision Language Models (VLMs) have enhanced the deep semantic understanding of autonomous driving systems, they exhibit notable limitations in decision explainability when handling rare and long-tail traffic scenarios. In this paper, we propose VLR-Driver, a novel multi-modal Vision-Language-Reasoning (VLR) framework based on Chain of Thought (CoT) for embodied autonomous driving. The framework employs a spatiotemporal CoT reasoning approach to recursively analyze potential safety risks and driving intentions of other agents, thereby delivering an efficient and transparent decision-making process. Furthermore, we construct a multi-modal reasoning-decision dataset to support the advancement of hierarchical reasoning of VLMs in autonomous driving. Closed-loop experiments conducted in CARLA demonstrate that the VLR-Driver significantly outperforms state-of-the-art end-to-end methods. Notably, key metrics such as driving score improved by 17.5%, while the success rate improved by 22.2%, offering a more transparent, reliable, and secure solution for autonomous driving systems. The code, dataset, and demonstration video will be open-sourced.

1. Introduction

In recent years, the rapid progress of End-to-End (E2E) architectures [6, 7, 18, 49], Large Language Models (LLMs) [12, 41, 48], and embodied intelligence [26, 53, 55] has established these technologies as key enablers of innovation in autonomous driving. Especially, VLMs enriched with extensive pre-training knowledge exhibit strong spatial understanding and common-sense reasoning abilities. DriveVLM [35] leverages VLM to enhance spatial awareness and planning capabilities in complex driving scenarios. CoVLA [2] integrates visual perception, language understanding, and action planning, demonstrating remarkable effectiveness in describing traffic scenarios and generating executable control actions.

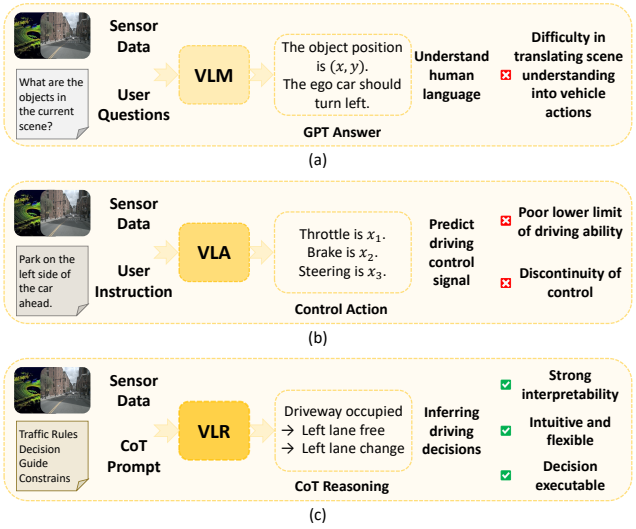


Figure 1. Comparison of different VLM-based AD systems. (a) VLM model focuses more on answering questions, but it is difficult to convert decisions into coherent control signals. (b) VLA model focuses more on predicting driving control signals, but lacks interpretability. (c) The VLR model generates decisions and controls signals with transparent reasoning processes through CoT, enhancing the driver trust in the system.

However, the decision-making process of VLMs often functions as a "black box", making it challenging to trace and interpret their underlying logic. This makes it difficult for autonomous driving systems to be fully trusted by drivers when encountering complex and emergency situations, such as illegal roadside parking, navigating intersec-

tions without traffic signals, and managing complex mixed-traffic interactions between motorized and non-motorized vehicles, thereby limiting their reliability and safety in real-world applications [39, 46, 54]. Moreover, most VLMs are trained on internet data, lacking spatial understanding and specialized training in the field of autonomous driving, making it difficult for them to fully adapt to dynamic and complex driving scenarios.

Meanwhile, CoT reasoning demonstrates strong inference, interpretability, and generalization capabilities by breaking down complex tasks into intermediate reasoning steps [40]. CoT enables systems to think step by step rather than relying on E2E black-box predictions, making it one of the key approaches toward achieving embodied autonomous driving. Sce2DriveX [52] enhances comprehensive perception and reasoning by introducing a multi-modal LLM framework with CoT, enabling a deeper understanding of spatiotemporal relationships and road topology. Similarly, DriveCoT [38] integrates CoT reasoning to improve decision-making interpretability and controll ability in autonomous driving systems. However, existing CoT-based methods heavily depend on predefined reasoning templates or limited training data, which may lead to misguided decisions in complex traffic scenarios. Additionally, current mathod primarily operate on static snapshots rather than continuous temporal sequences, limiting their ability to predict future events in dynamic traffic environments.

To bridge these gaps, in this work, we introduce **VLR-Driver**, a hierarchical CoT-based visual-language-reasoning model designed for closed-loop embodied autonomous driving. Our approach integrates the spatiotemporal features from cross-modal data, including multi-frame multi-view images and ego-vehicle control signals, by employing a SpatioTemporal CoT (ST-CoT) strategy that produces human-like reflective reasoning processes and driving action decisions. Additionally, we adopt a dual-phase training strategy, combining Low-Rank Adaptation (LoRA) [16] with an improved Stepwise Group Relative Policy Optimization (Step-GRPO) [34], significantly enhancing memory capacity and deep reasoning abilities. Our proposed VLR-Driver not only inherits the global action optimization capabilities of VLA models, but also preserves the transparency of modular rule-based methods. When encountering long-tail events and rare traffic scenarios, it demonstrates exceptional reflective reasoning and step-by-step inference processes, thereby enhancing human drivers' trust in autonomous driving systems. The differences between VLM, VLA, and our proposed VLR are illustrated in Fig. 1

To further enhance VLR models in environmental understanding, reasoning, and decision-making, we introduce the VLR-Driver Dataset. This data set includes detailed scene descriptions, weather information, vehicle state details, and most critically, human-like CoT reasoning processes and the corresponding driving decisions. Various experiments demonstrate that VLR-Driver is capable of making accurate driving decisions and coherent reasoning, even under challenging and highly dynamic road conditions.

The primary contributions of this work are summarized as follows:

- **Distinctive VLR-Driver Framework.** We introduce VLR-Driver, a visual-language-reasoning model developed for embodied autonomous driving. It generates a human-like reflective reasoning process within the driving system, enabling accurate driving decisions.
- **Spatiotemporal CoT.** We present a spatiotemporal CoT strategy that recursively analyzes potential safety risks and the driving intentions of moving agents in complex traffic scenarios, ensuring that the reasoning process of driving decisions remains transparent and interpretable.
- **Advanced VLR-Driver Dataset.** We construct VLR-Driver Dataset, a cutting-edge visual-language-reasoning and decision-making dataset specifically designed for autonomous driving. It supports the enhancement of spatiotemporal understanding and reflective reasoning capabilities in embodied autonomous driving systems.
- **Superior Performance in Closed-Loop Simulations.** Extensive closed-loop experiments are conducted on the CARLA platform. On the Bench2Drive benchmark, our approach achieves a 17.5% improvement in driving score and a 22.2% increase in success rate, providing a more human-like, reliable, and trustworthy solution.

2. Related Work

2.1. End-to-end Autonomous Driving

The rapid advancement of E2E-AD has fostered a growing transition away from modular rule-based methods toward data-driven approaches [15, 23]. Based on different input modalities, E2E methods can be categorized into visual-only methods [5, 8, 10, 17] and vision-LiDAR fusion methods [1, 20, 21]. TCP [42] and NEAT [10] adopt imitation learning methods, training directly on collected state-action pair datasets, demonstrating the feasibility of E2E approaches for autonomous driving. Roach [51] utilizes reinforcement learning experts as coaches, delivering dense and informative supervision signals to agents equipped with monocular camera inputs. However, vision-based methods inherently struggle with distance and depth estimation. These limitations may compromise the reliability of driving decisions [25]. To address these limitations, vision-LiDAR fusion methods such as TransFuser [11], CrossFuser [43], and FusionAD [47] have been designed to effectively integrate image with LiDAR data, significantly improving the robustness of AD systems in complex environments. However, different modalities contribute differently to the driving task, and spatiotemporal synchronization issues be-

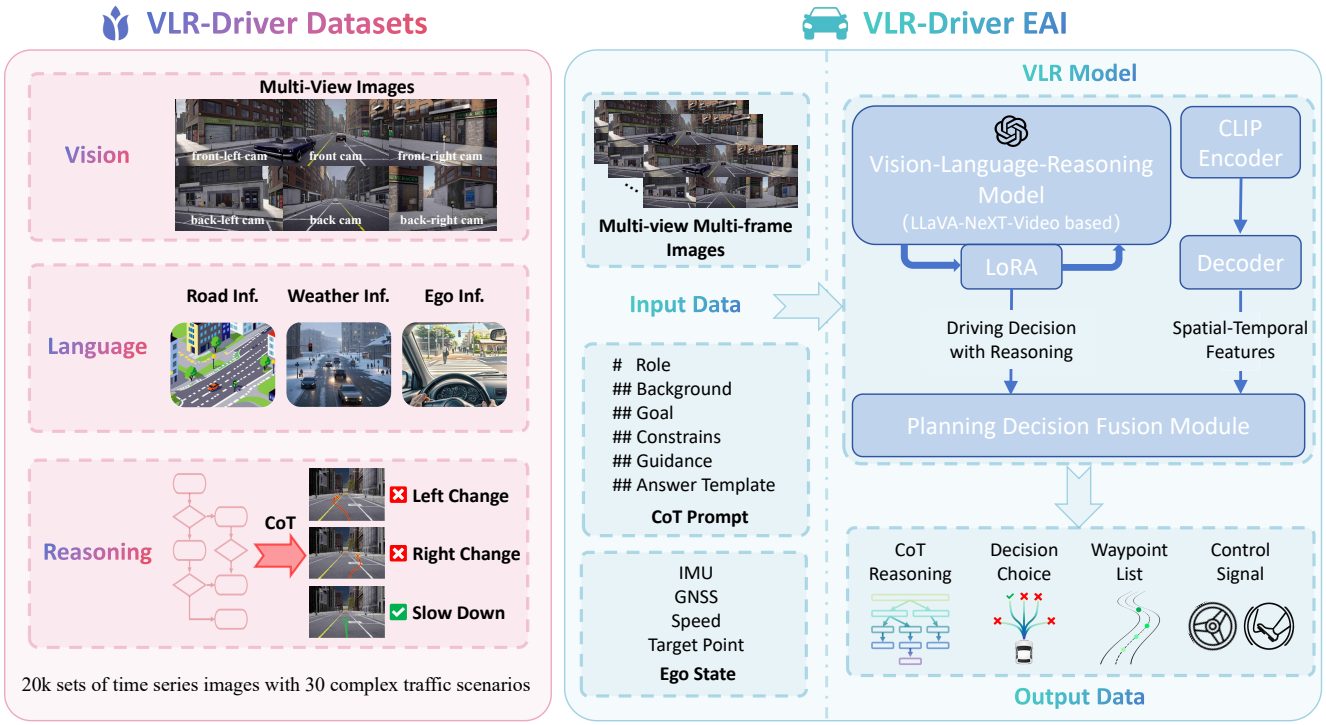


Figure 2. **Overview of VLR-Driver framework.** We introduce **VLR-Driver Dataset**, an advanced visual-language-reasoning dataset designed for autonomous driving, featuring detailed annotations of scene descriptions, analytical reasoning, and behavioral decisions. We present **VLR-Driver**, a novel multi-modal visual-language-reasoning framework for embodied autonomous driving that leverages a hierarchical spatiotemporal CoT reasoning mechanism.

tween sensor data can lead to inconsistencies, increasing both model complexity and training difficulty.

2.2. VLM and VLA in Autonomous Driving

VLMs unify visual perception with natural language processing capabilities, enabling a more comprehensive understanding of driving environments [19, 36, 37]. By incorporating cross-modal data fusion, integrating textual information alongside visual inputs, these systems gain text comprehension and human interaction capabilities that conventional E2E models inherently lack. DriveGPT4 [45] enhances LLMs' ability to process multimodal inputs by projecting them into the text domain, thereby enabling interpretable end-to-end autonomous driving. Senna [24] and DriveVLM [35] integrate VLMs with either traditional modular pipelines or E2E frameworks, achieving a decoupling between high-level planning and low-level trajectory prediction. This approach enhances planning performance while preserving the model's common sense reasoning capabilities.

Moreover, VLA represents an emerging paradigm that unifies visual perception, natural language understanding, and action prediction within a cohesive framework [13, 26]. Originally introduced in the field of robotics, RT-2 [4] pioneered the representation of robotic actions as text tokens, seamlessly incorporating them alongside natural language labels into the model's training set. This approach facilitates the direct transfer of internet-scale knowledge to robotic control, significantly enhancing both the generalization and semantic reasoning capabilities of robotic systems. In the context of autonomous driving, CoVLA [2] introduces an interpretable VLA model, seamlessly integrating visual perception, language-based scene understanding, and action planning. This integration enhances the system's ability to comprehend complex driving scenarios, anticipate trajectory outcomes, and execute informed driving decisions. However, despite these advancements, VLA models still exhibit limitations in accurately predicting precise control action values, leading to a low lower-bound in autonomous driving performance. Furthermore, when confronted with highly dynamic and complex traffic scenarios, their deep understanding and reasoning capabilities remain insufficient. Unlike the above methods, our proposed VLR model places greater emphasis on step-by-step reasoning and the thought process of the model, providing drivers with increased confidence in using autonomous driving systems.

2.3. Chain of Thought

The CoT technology is an extension of prompt engineering, proposed by Wei Jason in 2022 [40], which has greatly

improved the effectiveness of reasoning for complex problems. The highly anticipated Deepseek-R1 [33] model also utilizes the CoT technique, which deeply integrates multimodal knowledge base data, enabling the model to generate a step-by-step thinking process. DriveCoT [38] has built a CoT dataset that includes sensor data, control decisions, and CoT labels used to indicate reasoning processes. The model trained on this basis can generate predictions and final decisions with CoT, effectively improving model performance. LanguageMPC [30] combines LLM with Model Predictive Control and decomposes driving decisions into multiple subtasks through a CoT framework. This method enables the auto drive system to think like human beings, and improves its ability to handle complex scenes. OpenEMMA [44] introduces CoT technology to guide model generation of detailed descriptions of key objects, behavioral insights, and meta driving decisions, improving system transparency and usability. Motivated by these advancements, we apply step-by-step hierarchical spatiotemporal CoT to autonomous driving, enhancing the interpretability of reasoning and decision-making.

3. Method

We present the motivation and design details of our VLR-Driver framework. As depicted in Fig. 2, VLR-Driver comprises two main components: the VLR-Driver Embodied Agent and the VLR-Driver Dataset. Initially, we introduce the design concept of the VLR model, which builds upon enhancements to the VLA model (Sec. 3.1). Subsequently, we elaborate on the hierarchical spatiotemporal CoT methodology (Sec. 3.2) and the specifics of the dual-phase training strategy (Sec. 3.3).

3.1. Overview

The VLR model is a large visual-language-reasoning model designed for embodied autonomous driving. It can processes visual inputs, such as multi-view images, alongside textual information, including vehicle control signals. The model is capable of extracting spatiotemporal key features within a multi-modal embedding and recursively analyzing potential safety risks and the driving intentions of other agents within the reasoning level of the VLR model. Ultimately, it formulates a comprehensive reasoning framework and well-structured decision-making outputs, explicitly identifying critical risk factors and the underlying rationale behind each decision. This process significantly reinforces the robustness and safety of the autonomous driving system, ensuring adaptive resilience in complex and dynamic environments.

In this study, we employ pre-trained LLaVA-NeXT-Video [50] as the VLM and CLIP [29] as the visual encoder. The model is capable of processing multi-frame multi-view image data that capture historical temporal context, while also extracting real-time vehicle sensor information to facilitate dynamic and context-aware decision-making.

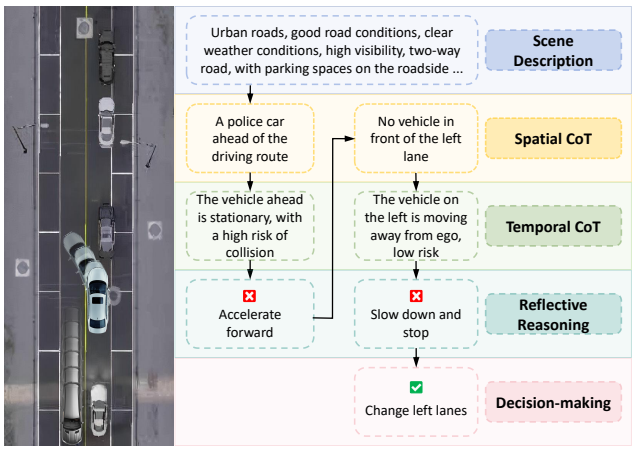


Figure 3. Illustration of the ST-CoT reasoning process. In this scenario, where some vehicles are illegally parked ahead and blocking the lane, our method can conduct hierarchical patiotemporal reasoning analysis and make a decision of change left lane once the adjacent lane is free.

Input Representations. We utilize N_f frame and N_v view images from the past period, with a field of view (FoV) of 70 degrees. It can be represented as $V \in \mathbb{R}^{N_f \times N_v \times 3 \times H_0 \times W_0}$. At the same time, there are also the current position (x, y) of ego vehicle, the speed v, the target point position (p, q). Subsequently, the compressed and cropped image data and the information from the ego's sensors are input into the model.

Output Representations. Our output consists of the reasoning process and driving decisions generated by the VLR model for the current driving scenario. This includes risk identification, traffic signal recognition, motion direction prediction, and autonomous driving decision-making. The driving decision will also be fused with the spatiotemporal information features extracted by the E2E model, and finally output the waypoints and control signals of the vehicle for the next moment.

3.2. Spatiotemporal CoT Reasoning

To enhance the reasoning capabilities and transparency of the autonomous driving system, we introduce a hierarchical ST-CoT that guides the model to approach driving decisions in a human-like manner. Our method decomposes the driving decision-making process into two levels: a perception-level spatiotemporal CoT $C_{perception}$, which focuses on extracting and understanding environmental dynamics, and a decision-level dynamic CoT $C_{decision}$, which refines and optimizes decision-making based on contextual and temporal factors. The example of CoT is shown in Fig. 3.

3.2.1. Perception Level CoT

The spatiotemporal characteristics of the environment play a crucial role in the autonomous driving process. The perception level CoT is responsible for guiding the model to extract spatial and temporal features from input image data, identify and locate crucial objects in the traffic, such as vehicles, pedestrians, alien objects, traffic lights, traffic signs, etc., and extract historical behavioral features of dynamic agent based on temporal information. Our method enables VLR model to describe the current driving scenario, construct real-time spatial layout and dynamic changes of the environment, and achieve long-term planning for driving decisions.

Spatial CoT. In driving scenarios, we primarily focus on obstacles that impact the ego's normal operation, including object categories O_{type} and distances with the ego O_{dis}. A critical aspect of safe driving is identifying potential risk points within the current lane. Additionally, when the vehicle executes lane changes, objects in adjacent lanes, both left and right, may significantly influence its movement. Beyond obstacles, key traffic light S_{light}, traffic signs S_{sign}, and lane markings S_{mark} are also integral to decision-making, ensuring comprehensive spatial awareness.

Temporal CoT. While a single-frame image can provide a static representation of road scenes and traffic participants, it fails to capture the motion trends of moving agents. To address this limitation, we introduce consecutive frames $I = \{I_f, I_{fr}, I_{fl}, I_b, I_{bl}, I_{br}\}_{t=T_{now}-T}^{T_{now}}$ into the model, allowing it to track temporal variations in object positioning, which I_f, I_{fr}, I_{fl}, I_b, I_{bl}, I_{br} represents the image view of front, front left, front right, back, back left and back right. These sequential frames not only offer instantaneous spatial context but also reveal motion trajectories and behavioral patterns through their inter-frame positional changes. This temporal information is essential for predicting dynamic object movement, assessing collision risks, and generating robust path planning strategies, ultimately enhancing the ability of anticipate and react to evolving traffic conditions.

3.2.2. Decision Level CoT

The output information from the perception-level serves as a critical foundation for the driving decision level, enabling reliable autonomous driving behavior inference. Within the driving decision level, we account for complex dynamic environmental factors, transforming spatial and temporal information from the perception level into concrete driving decisions. Specifically, decision-makers must not only analyze the current driving environment in real time but also anticipate future behaviors of other traffic participants and make decisions based on multiple factors. Throughout this process, the CoT in the driving decision level spans multiple perspectives, incorporating safety, efficiency, comfort, and compliance with traffic regulations to ensure that driving decisions meet safety standards while optimizing driving efficiency.

To further enhance the structured reasoning process in autonomous driving, we have carefully designed hierarchical reasoning prompts that guide decision-making. Our structured prompts follow a logical sequence of "risk point recognition — driving intention prediction — driving decision selection", forming a cohesive reasoning chain aligned with human cognitive driving patterns.

Risk Point Recognition. In this initial stage, the prompt-driven model conducts a comprehensive perception and analysis of the driving environment. This includes recognizing and evaluating critical elements such as traffic signs, lane markings, pedestrians, and obstacles to identify potential risks.

Driving Intention Prediction. Once risk points are identified, the model leverages dynamic target behavior prediction and scene understanding to infer the potential movements and intentions of other road users. For instance, the model assesses whether pedestrians are likely to cross the road or whether the vehicle ahead intends to change lanes.

Driving Decision Selection. Based on the contextual information gathered from the first two stages, the model applies multimodal information fusion and weighted decision-making to select the most optimal driving maneuver.

Through this structured prompting strategy, the large model adheres to a progressive reasoning hierarchy, beginning with fundamental environmental perception and advancing to higher-level decision-making. By explicitly presenting the reasoning process in a clear and structured manner, this approach enhances passenger trust and confidence in the intelligent driving system. Moreover, the integration of recursive CoT reasoning enables the model to mimic the step-by-step thought process of human drivers, facilitating more flexible, reliable, and interpretable decision outputs in complex driving scenarios. The structured prompt framework are shown in Fig. 3.

3.3. Training Paradigm

VLR-Driver adopts a dual-phase training strategy to optimize its reasoning and decision-making capabilities. In the first phase, LoRA [16] is utilized for supervised fine-tuning on a pre-trained large model, enabling efficient adaptation with minimal memory and computational overhead while maintaining strong performance. In the second phase, Step-GRPO is applied for reinforcement learning based on human preferences, further enhancing the model's ability to

exhibit human-like reasoning and decision-making characteristics.

Training with LoRA. LoRA is a parameter-efficient fine-tuning technique that enables effective model adaptation by performing a low-rank decomposition of the weight matrix. The core principle behind LoRA is to decompose the weight matrix of a pre-trained model into a low-rank structure, significantly reducing the number of trainable parameters while preserving expressivity. Specifically, if the weight matrix in the pre-trained model is denoted as $\boldsymbol{W_0} \in \mathbb{R}^{d\times k}$, LoRA represents it as:

$$\boldsymbol{W'} = \boldsymbol{W_0} + \Delta \boldsymbol{W} = \boldsymbol{W_0} + \frac{\alpha}{r}\boldsymbol{B}\cdot\boldsymbol{A}, \tag{1}$$

where $\boldsymbol{A} \in \mathbb{R}^{r\times k}$ and $\boldsymbol{B} \in \mathbb{R}^{d\times r}$, r is the rank of a low rank matrix, usually much smaller than d and k, α is a scaling factor. The forward pass is computed as:

$$y = \boldsymbol{W'}x = \left(\boldsymbol{W_0} + \frac{\alpha}{r}\boldsymbol{B}\cdot\boldsymbol{A}\right)x, \tag{2}$$

where y is the output and x is input.

We use LoRA for all linear modules, which not only saves computation but also ensures the performance of the model.

Training with GRPO. The core principle of GRPO [34] is to optimize strategies by assigning relative rewards to multiple outputs generated from the same prompt, thereby eliminating the need for additional value function models. The introduction of process reward model estimation in GRPO provides finer support for distributed rewards. The reward of each step of the outputs is:

$$R = \left\{\left\{r_1^{index(1)}, \cdots, r_1^{index(K_1)}\right\}, \cdots, \left\{r_G^{index(1)}, \cdots, r_G^{index(K_G)}\right\}\right\}, \tag{3}$$

where $index(\cdot)$ is the end token index, and the reward need be normalized as:

$$\widetilde{r}_i^{index(j)} = \frac{r_i^{index\,(j)} - \text{mean}(\boldsymbol{R})}{\text{std}(\boldsymbol{R})}. \tag{4}$$

We extend GRPO by introducing reasoning Step-GRPO, a supervised reasoning decision process that structures the output into multiple steps based on the CoT reasoning framework. At each step, a reward function is applied to evaluate and assign scores, enabling fine-grained feedback that enhances model interpretability and accelerates convergence. Specifically, we first generate multiple candidate decision answers for the current driving scenario using prompts within the VLR model; Then, following our ST-CoT strategy, the reasoning process is divided into four distinct steps: scene description, spatial risk point reasoning, dynamic trajectory prediction, and driving decision-making; Furthermore, each reasoning step is assigned a reward to encourage structured learning. Finally, we compare all answers within the group and calculate the Kullback-Leibler (KL) divergence to update the policy model. This grouping and step-by-step scoring strategy enhances training efficiency and reduces the likelihood of erroneous reasoning in the model.

4. VLR-Driver Dataset

To fully explore the reasoning and decision-making capabilities of large language models, we propose an advanced reasoning and decision-making dataset for autonomous driving scenarios, called the VLR-Driver Datasets. This dataset relies on the CARLA [14] simulator for data collection and is expanded and meticulously annotated based on the Bench2Drive dataset [22]. It includes: a) multi-view multi-frame images or videos, b) valuable information for autonomous driving, such as road details, weather conditions, vehicle information, and scene descriptions, and c) driving decision choices along with the decision-making process based on a ST-CoT. This dataset provides a rich and comprehensive training foundation for autonomous driving reasoning and decision-making, allowing the agent to exhibit human-like reasoning while interacting with the environment. The dataset includes 20,000 sets of multi-frame, multi-angle image data collected from various road conditions such as urban, rural, and highways in the CARLA simulator, covering over 40 specific complex traffic scenarios (e.g., forward accidents, dynamic object crossings, etc.). Each image set provides road scene descriptions, environmental weather information, vehicle status data, and, most importantly, the human-like reasoning process and driving behavior decisions.

4.1. Data Collection

We conducted data collection based on the 44 corner scene classifications provided by Bench2Drive to ensure optimal autonomous driving performance in various complex corner scenarios. The Bench2Drive dataset offers a rich array of data and annotations, including multi-angle images, lidar, radar, vehicle information, and expert assessments, which have been instrumental in building our VLR dataset. Additionally, we selected over 40 scenes where autonomous driving's reasoning and decision-making capabilities are relatively weak, using them as the visual and textual components of our dataset. We further expanded and enriched the dataset by collecting additional data in the CARLA simulator. We have taken into account various weather conditions, road conditions, and the types and numbers of traffic participants in the scene.

Weather. For each scene, we randomly set values for cloudiness, fog density, precipitation, precipitation deposits, sun altitude angle, sun azimuth angle, wetness, and wind intensity. The combinations of these parameters cover a variety of weather conditions, such as sunny, rainy, foggy,

Table 1. The comparison of core metrics and subdivision infraction scores with state-of-the-art E2E/VLM models on the Bench2Drive benchmark. C, L and T indicate camera, LiDAR and text modalities, respectively. DS, RC, IS, SR correspond to the Driving Score, Route Completion, Infraction Score, and Success Rate. CP, CV, CL, RL, SS, OR, AB, YEV correspond to the Collision with a Pedestrian, Collision with another Vehicle, Collision with Layout, Red Light infractions, Stop Sign infractions, Off-Road infractions, Agent Blocked, and failure to Yield to Emergency Vehicles infractions.

Method	Type	Modality	Core Metrics ↑				Subdivision Infraction Score ↓							
			DS	RC	IS	SR	CP	CV	CL	RL	SS	OR	AB	YEV
NEAT [10]	E2E	C	30.86	55.35	0.55	6.81	1.08	9.87	5.57	0.20	1.33	0.41	2.01	0.27
TCP [42]		C	56.28	83.57	0.65	25.00	**0.26**	5.46	5.46	**0.00**	0.52	0.22	0.78	**0.00**
LeTFuser [1]		C+L	52.53	77.68	0.67	18.18	1.16	5.54	3.79	0.29	0.87	0.12	0.58	0.29
LateFusion [28]		C+L	48.53	58.32	0.85	18.18	0.38	3.11	1.55	0.38	0.77	0.06	1.55	0.38
TransFuser [11]		C+L	37.18	68.14	0.51	9.09	0.96	13.24	8.71	**0.00**	0.96	0.32	2.58	0.32
ThinkTwice [21]		C+L	58.79	74.35	0.77	29.54	0.30	5.76	0.91	**0.00**	0.91	0.05	0.91	0.30
EATNet [9]		C+L	42.97	78.84	0.54	15.91	0.82	14.01	1.92	**0.00**	1.64	0.22	1.09	0.27
InterFuser [31]		C+L	63.81	80.46	0.79	40.90	0.35	3.81	0.54	0.27	1.08	0.05	0.54	0.27
LMDrive [32]	VLM	C+L+T	24.76	33.02	**0.90**	13.63	1.14	2.86	2.29	**0.00**	0.57	0.05	3.44	0.57
LeapAD [27]		C+T	55.18	77.45	0.71	36.36	0.69	5.07	1.15	0.20	0.91	0.08	1.47	0.27
VLR-Driver (Ours)	VLR	C+T	**75.00**	**86.08**	0.87	**50.00**	0.72	**2.83**	**0.48**	**0.00**	**0.48**	**0.04**	**0.24**	0.24

Table 2. The comparison of driving advanced ability and experience score with state-of-the-art models on the Bench2Drive benchmark. OT, MER, EB, GW, TS, DE, SC correspond to the OverTaking, MERging, Emergency Brake, Give Way, Traffic Sign, Driving Efficiency, and Smoothness Control.

Method	Driving Advanced Ability ↑						Exper. Score ↑	
	OT	MER	EB	GW	TS	Mean	DE	SC
NEAT [10]	0.00	6.66	9.09	**50.00**	27.77	18.70	92.09	0.30
TCP [42]	25.00	13.33	27.27	**50.00**	50.00	33.12	114.52	0.27
LeTFuser [1]	0.00	20.00	18.18	0.00	47.22	17.08	115.09	0.47
LateFusion [28]	0.00	20.00	9.09	0.00	36.11	13.04	104.39	**0.59**
TransFuser [11]	0.00	13.33	9.09	0.00	36.11	11.71	95.20	0.36
ThinkTwice [21]	12.50	20.00	36.36	**50.00**	52.77	34.33	91.17	0.36
EATNet [9]	0.00	13.33	18.18	0.00	41.66	14.63	88.13	0.33
InterFuser [31]	0.00	46.66	54.54	**50.00**	61.11	42.46	119.20	0.32
LMDrive [32]	25.00	6.66	9.09	**50.00**	2.77	18.70	75.41	0.22
LeapAD [27]	12.50	33.33	27.27	**50.00**	44.44	33.51	93.33	0.26
VLR-Driver (Ours)	**37.50**	**46.66**	**72.72**	**50.00**	**72.22**	**55.82**	**125.22**	**0.59**

broken sky, and stormy, as well as different lighting conditions for day and night.

Roads. The scenes include various road types such as urban two-way single-lane roads, multi-lane roads, highways, and narrow rural roads.

Traffic Participants. Different corner cases involve various traffic participants, including cars, bicycles, pedestrians, ambulances, etc., simulating the complex traffic conditions encountered in daily driving.

4.2. Data Annotation

The reasoning chain process and the decision-making choices are crucial for training large autonomous driving models and are key to enhancing the model's human-like reasoning ability. Therefore, we performed secondary annotation on the dataset we collected.

Driving Scene Descriptions. We used the pre-trained large visual language model Qwen2-VL [3] to generate detailed descriptions for the corresponding driving scenes. These descriptions primarily focus on environmental information such as road conditions, weather, and lighting, as well as dynamic targets like vehicles and pedestrians that may pose driving risks.

Reasoning Decisions and Process. We considered pre-annotated information such as vehicle speed, acceleration, steering angle, traffic light status, and the state of the vehicle ahead. A rule-based method was used to determine the true values of future motion behaviors based on the decision choices made at earlier time steps. Additionally, we completed the predefined CoT reasoning text statements. Finally, to ensure the accuracy and consistency of the annotations, especially for decision choices, carefully manual verification was carried out.

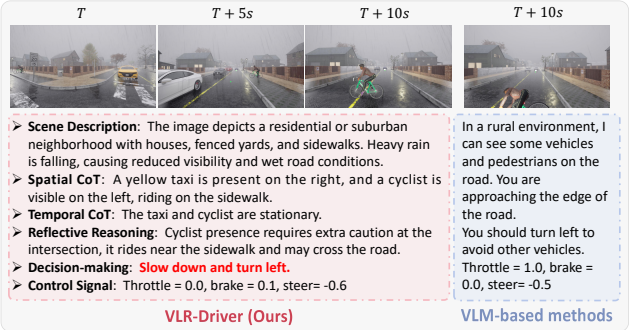
5. Experiment

5.1. Experimental Setup

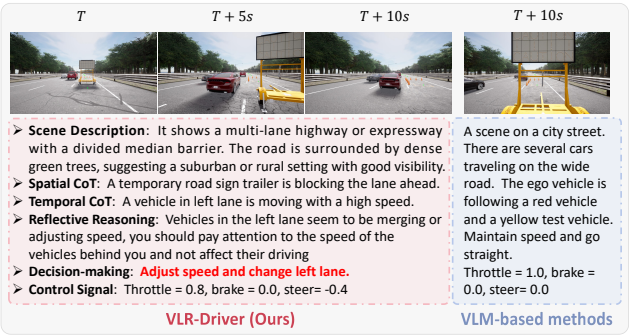
Our method was validated on the open source autonomous driving simulation platform CARLA 0.9.15 [14]. The VLR-Driver model was trained on a server equipped with 8 NVIDIA A800 GPUs (each with 80G of video memory) for approximately 50 hours. The dataset used was the VLR-Driver dataset that we developed. Specifically, we use ViT-g/14 from EVA-CLIP [29] as the vision encoder and LLaVA-NeXT-Video-7B [50] as the VLM. The resolution of the input image is set to 336 $\times$ 336 pixels.

5.2. Metrics

We employ four core metrics to evaluate autonomous driving performance: driving score (DS), route completion (RC), infraction score (IS), and success rate (SR). Additionally, to provide a more granular assessment of model performance in specific aspects, we introduce three supplementary evaluation categories: subdivision infraction score, driving advanced ability, and driving experience score [22].



(a) Vehicle turning route with bicycle crossing.



(b) Construction obstacle.

Figure 4. Visual comparison between VLR-Driver and VLM-based methods. The ST-CoT guides the VLR model to approach driving decisions in a human-like spatiotemporal manner. Based on the sequence of images from the preceding time period T, we derive the following inference results. The images captured at 5 and 10 seconds afterward validate the accuracy of our decisions.

5.3. Comparisons with Existing Methods

We conducted comprehensive experiments with the SOTA methods including E2E and VLM in the CARLA simulator with Bench2Drive Benchmark. We present comparison result in Tab. 1. It can be seen that our method outperforms other methods in key metrics such as DS, RC, and SR, achieving first place and effectively improving DS by 17.5%, mean of driving ability by 31.4%, and SR by 22.2%.

The comparison results of the driving advanced ability and driving experience score of each method are shown in Tab. 2. Our method achieved the best results in all abilities, thanks to the deep reflection and reasoning ability of our VLR model, which has stronger traffic reasoning capacity in special road conditions. Most E2E methods can only achieve following the vehicle, but when there is a vehicle temporarily parked in the lane ahead, blocking the self driving route, they will keep stopping and waiting, making it impossible to complete the entire route. And our VLR-Driver can achieve deep understanding and inference of the current scene through large-scale model inference, so as to make timely detours.

Table 3. Ablation study for each module.

ID	Abal.	Exp.	Core Metrics ↑		Driving Advanced Ability ↑		
			DS	SR	OT	MER	EB
1	VLR	Full Model	**71.48**	**54.54**	**37.50**	**46.66**	**72.72**
2	Arch.	w/o CoT	57.17	34.09	25.00	40.00	41.66
3		w/o VLR-Model	46.39	20.45	14.28	26.66	36.36
4	Train	w/o VLR-Data	52.85	27.27	25.00	26.66	9.09
5		w/o Step-GRPO	65.57	45.45	37.50	40.00	63.63

5.4. Ablation Study

We conducted a comprehensive ablation study, detailed in Tab. 3. The experimental configurations include four variants: (1) Without utilizing our proposed spatiotemporal CoT strategy, using only a question-based approach without reasoning guidance. (2) Without using our VLR model, instead employing a standard LLM module. (3) Removing the VLR data used to guide the reasoning process. (4) Without Step-GRPO reinforcement learning training, using only LoRA strategies to fine-tune the model with supervision. The results show the effectiveness of each contribution.

5.5. Visualization

We selected some special scenarios to visualize the performance in complex traffic situations with reflective reasoning, and the comparison results shown in Fig. 4. The ST-CoT enables the model to make driving decisions with human-like reasoning, considering both spatial and temporal dynamics. More visual comparison results can be found in the Appendix.

6. Conclusion

In this paper, we introduce VLR-Driver, a VLR model for embodied AD. It leverages a carefully designed ST-CoT strategy to guide the model in recursively analyzing potential safety risks and the driving intentions of dynamic agents in complex traffic scenarios. Our dual-phase training method significantly enhances the generalization of the model. Additionally, we propose the VLR-Driver dataset, which effectively integrates spatiotemporal perception, language understanding, and reflective reasoning, providing crucial support for interpretability reasoning in AD system. Experimental results show that VLR-Driver outperforms other methods on Bench2Drive, achieving cutting-edge performance and paving the way for EAI realization.

Limitations and future work. There are differences between the data in simulation platforms and the real world. How to transfer and adapt it to a real-world style remains an important area for further exploration.

References

[1] Pedram Agand, Mohammad Mahdavian, Manolis Savva, and Mo Chen. Letfuser: Light-weight end-to-end transformer-based sensor fusion for autonomous driving with multi-task learning. 2023. 2, 7

[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*, abs/2408.10845, 2024. 1, 3

[3] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023. 7

[4] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding, Danny Driess, Avinava Dubey, Chelsea Finn, Pete Florence, Chuyuan Fu, Montse Gonzalez Arenas, Keerthana Gopalakrishnan, Kehang Han, Karol Hausman, Alexander Herzog, Jasmine Hsu, Brian Ichter, Alex Irpan, Nikhil Joshi, Ryan Julian, Dmitry Kalashnikov, Yuheng Kuang, Isabel Leal, Lisa Lee, Tsang-Wei Edward Lee, Sergey Levine, Yao Lu, Henryk Michalewski, Igor Mordatch, Karl Pertsch, Kanishka Rao, Krista Reymann, Michael Ryoo, Grecia Salazar, Pannag Sanketi, Pierre Sermanet, Jaspiar Singh, Anikait Singh, Radu Soricut, Huong Tran, Vincent Vanhoucke, Quan Vuong, Ayzaan Wahid, Stefan Welker, Paul Wohlhart, Jialin Wu, Fei Xia, Ted Xiao, Peng Xu, Sichun Xu, Tianhe Yu, and Brianna Zitkovich. Rt-2: Vision-language-action models transfer web knowledge to robotic control. 2023. 3

[5] Daoming Chen, Ning Wang, Feng Chen, and Tony Pipe. Detrive: Imitation learning with transformer detection for end-to-end autonomous driving. 2023. 2

[6] Jianyu Chen, Zhuo Xu, and Masayoshi Tomizuka. End-to-end autonomous driving perception with sequential latent representation learning. 2020. 1

[7] Li Chen, Penghao Wu, Kashyap Chitta, Bernhard Jaeger, Andreas Geiger, and Hongyang Li. End-to-end autonomous driving: Challenges and frontiers. 2023. 1

[8] Shaoyu Chen, Bo Jiang, Hao Gao, Bencheng Liao, Qing Xu, Qian Zhang, Chang Huang, Wenyu Liu, and Xinggang Wang. Vadv2: End-to-end vectorized autonomous driving via probabilistic planning. 2024. 2

[9] Weihuang Chen, Fanjie Kong, Liming Chen, Shen'ao Wang, Zhiping Wang, and Hongbin Sun. Eatnet: Efficient axial transformer network for end-to-end autonomous driving. In *ITSC*, 2024. 7

[10] Kashyap Chitta, Aditya Prakash, and Andreas Geiger. *NEAT: Neural Attention Fields for End-to-End Autonomous Driving*. 2021. 2, 7

[11] Kashyap Chitta, Aditya Prakash, Bernhard Jaeger, Zehao Yu, Katrin Renz, and Andreas Geiger. Transfuser: Imitation with transformer-based sensor fusion for autonomous driving. *PAMI*, Vol.45(No.11):12878–12895, 2023. 2, 7

[12] Can Cui, Yunsheng Ma, Xu Cao, Wenqian Ye, Yang Zhou, Kaizhao Liang, Jintai Chen, Juanwu Lu, Zichong Yang, Kuei-Da Liao, Tianren Gao, Erlong Li, Kun Tang, Zhipeng Cao, Tong Zhou, Ao Liu, Xinrui Yan, Shuqi Mei, Jianguo Cao, Ziran Wang, and Chao Zheng. A survey on multimodal large language models for autonomous driving. 2023. 1

[13] Pengxiang Ding, Han Zhao, Zhitao Wang, Zhenyu Wei, Shangke Lyu, and Donglin Wang. Quar-vla: Vision-language-action model for quadruped robots. 2023. 3

[14] Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. Carla: An open urban driving simulator. In *CoRL*, pages 1–16. PMLR, 2017. 6, 7

[15] Hao Gao, Shaoyu Chen, Bo Jiang, Bencheng Liao, Yiang Shi, Xiaoyang Guo, Yuechuan Pu, Haoran Yin, Xiangyu Li, Xinbang Zhang, Ying Zhang, Wenyu Liu, Qian Zhang, and Xinggang Wang. Rad: Training an end-to-end driving policy via large-scale 3dgs-based reinforcement learning. 2

[16] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022. 2, 5

[17] 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*. 2022. 2

[18] Zhiyu Huang, Haochen Liu, and Chen Lv. Gameformer: Game-theoretic modeling and learning of transformer-based interactive prediction and planning for autonomous driving. In *ICCV*, pages 3903–3913. 1

[19] Jyh-Jing Hwang, Runsheng Xu, Hubert Lin, Wei-Chih Hung, Jingwei Ji, Kristy Choi, Di Huang, Tong He, Paul Covington, and Benjamin Sapp. Emma: End-to-end multimodal model for autonomous driving. *arXiv preprint arXiv:2410.23262*, 2024. 3

[20] Xiaosong Jia, Yulu Gao, Li Chen, Junchi Yan, Patrick Langechuan Liu, and Hongyang Li. Driveadapter: Breaking the coupling barrier of perception and planning in end-to-end autonomous driving. In *ICCV*, pages 7953–7963. 2

[21] Xiaosong Jia, Penghao Wu, Li Chen, Jiangwei Xie, Conghui He, Junchi Yan, and Hongyang Li. *Think Twice before Driving: Towards Scalable Decoders for End-to-End Autonomous Driving*. 2023. 2, 7

[22] Xiaosong Jia, Zhenjie Yang, Qifeng Li, Zhiyuan Zhang, and Junchi Yan. Bench2drive: Towards multi-ability benchmarking of closed-loop end-to-end autonomous driving. *arXiv preprint arXiv:2406.03877*, 2024. 6, 7

[23] Xiaosong Jia, Junqi You, Zhiyuan Zhang, and Junchi Yan. Drivetransformer: Unified transformer for scalable end-to-end autonomous driving. In *ICLR*, 2025. 2

[24] Bo Jiang, Shaoyu Chen, Bencheng Liao, Xingyu Zhang, Wei Yin, Qian Zhang, Chang Huang, Wenyu Liu, and Xinggang Wang. Senna: Bridging large vision-language models and end-to-end autonomous driving. . 3

[25] 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 *ICCV*, pages 8340–8350, . 2

[26] Yueen Ma, Zixing Song, Yuzheng Zhuang, Jianye Hao, and Irwin King. A survey on vision-language-action models for embodied ai. *ArXiv*, abs/2405.14093, 2024. 1, 3

[27] Jianbiao Mei, Yukai Ma, Xuemeng Yang, Licheng Wen, Xinyu Cai, Xin Li, Daocheng Fu, Bo Zhang, Pinlong Cai, and Min Dou. Continuously learning, adapting, and improving: A dual-process approach to autonomous driving. *arXiv preprint arXiv:2405.15324*, 2024. 7

[28] Aditya Prakash, Kashyap Chitta, and Andreas Geiger. *Multi-Modal Fusion Transformer for End-to-End Autonomous Driving*. 2021. 7

[29] 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 *ICML*, pages 8748–8763. PmLR, 2021. 4, 7

[30] Hao Sha, Yao Mu, Yuxuan Jiang, Li Chen, Chenfeng Xu, Ping Luo, Shengbo Eben Li, Masayoshi Tomizuka, Wei Zhan, and Mingyu Ding. Languagempc: Large language models as decision makers for autonomous driving. *arXiv preprint arXiv:2310.03026*, 2023. 4

[31] Hao Shao, Letian Wang, RuoBing Chen, Hongsheng Li, and Yu Liu. Safety-enhanced autonomous driving using interpretable sensor fusion transformer. 2022. 7

[32] Hao Shao, Yuxuan Hu, Letian Wang, Steven L. Waslander, Yu Liu, and Hongsheng Li. Lmdrive: Closed-loop end-to-end driving with large language models. *CVPR*, 2024. 7

[33] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Y Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024. 4

[34] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Y Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024. 2, 6

[35] 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. 2024. 1, 3

[36] Junming Wang, Xingyu Zhang, Zebin Xing, Songen Gu, Xiaoyang Guo, Yang Hu, Ziying Song, Qian Zhang, Xiaoxiao Long, and Wei Yin. He-drive: Human-like end-to-end driving with vision language models. *arXiv preprint arXiv:2410.05051*, 2024. 3

[37] Shihao Wang, Zhiding Yu, Xiaohui Jiang, Shiyi Lan, Min Shi, Nadine Chang, Jan Kautz, Ying Li, and Jose M. Alvarez. Omnidrive: A holistic llm-agent framework for autonomous driving with 3d perception, reasoning and planning. 2024. 3

[38] Tianqi Wang, Enze Xie, Ruihang Chu, Zhenguo Li, and Ping Luo. Drivecot: Integrating chain-of-thought reasoning with end-to-end driving. *ArXiv*, abs/2403.16996, 2024. 2, 4

[39] Wenhai Wang, Jiangwei Xie, ChuanYang Hu, Haoming Zou, Jianan Fan, Wenwen Tong, Yang Wen, Silei Wu, Hanming Deng, Zhiqi Li, Hao Tian, Lewei Lu, Xizhou Zhu, Xiaogang Wang, Yu Qiao, and Jifeng Dai. Drivemlm: Aligning multi-modal large language models with behavioral planning states for autonomous driving. 2023. 2

[40] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. In *NeurIPS*, pages 24824–24837, 2022. 2, 3

[41] Jiayang Wu, Wensheng Gan, Zefeng Chen, Shicheng Wan, and Philip S. Yu. Multimodal large language models: A survey. 2023. 1

[42] 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. 2022. 2, 7

[43] Weishang Wu, Xiaoheng Deng, Ping Jiang, Shaohua Wan, and Yuanxiong Guo. Crossfuser: Multi-modal feature fusion for end-to-end autonomous driving under unseen weather conditions. *ITSC*, Vol.24(No.12):1–15, 2023. 2

[44] Shuo Xing, Chengyuan Qian, Yuping Wang, Hongyuan Hua, Kexin Tian, Yang Zhou, and Zhengzhong Tu. Openemma: Open-source multimodal model for end-to-end autonomous driving. *ArXiv*, abs/2412.15208, 2024. 4

[45] Zhenhua Xu, Yujia Zhang, Enze Xie, Zhen Zhao, Yong Guo, Kenneth K.Y. Wong, Zhenguo Li, and Hengshuang Zhao. Drivegpt4: Interpretable end-to-end autonomous driving via large language model. 2023. 3

[46] Zhenjie Yang, Xiaosong Jia, Hongyang Li, and Junchi Yan. Llm4drive: A survey of large language models for autonomous driving. 2023. 2

[47] Tengju Ye, Wei Jing, Chunyong Hu, Shikun Huang, Lingping Gao, Fangzhen Li, Jingke Wang, Ke Guo, Wencong Xiao, Weibo Mao, Hang Zheng, Kun Li, Junbo Chen, and Kaicheng Yu. Fusionad: Multi-modality fusion for prediction and planning tasks of autonomous driving. 2023. 2

[48] Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. A survey on multimodal large language models. 2023. 1

[49] Jiakai Zhang. *End-to-end Learning for Autonomous Driving*. Thesis, 2019. 1

[50] Yuanhan Zhang, Bo Li, haotian Liu, Yong jae Lee, Liangke Gui, Di Fu, Jiashi Feng, Ziwei Liu, and Chunyuan Li. Llava-next: A strong zero-shot video understanding model, 2024. 4, 7

[51] Zhejun Zhang, Alexander Liniger, Dengxin Dai, Fisher Yu, and Luc Van Gool. End-to-end urban driving by imitating a reinforcement learning coach. 2021. 2

[52] Rui Zhao, Qirui Yuan, Jinyu Li, Haofeng Hu, Yun Li, Chengyuan Zheng, and Fei Gao. Sce2drivex: A generalized mllm framework for scene-to-drive learning. *arXiv preprint arXiv:2502.14917*, 2025. 2

[53] Z. Zhao, Q. Wu, J. Wang, B. Zhang, C. Zhong, and A. A. Zhilenkov. Exploring embodied intelligence in soft robotics: A review. *Biomimetics (Basel)*, 9(4), 2024. 1

[54] Xingcheng Zhou, Mingyu Liu, Ekim Yurtsever, Bare Luka Zagar, Walter Zimmer, Hu Cao, and Alois C. Knoll. Vision language models in autonomous driving: A survey and outlook. *TIV*, pages 1–20, 2024. 2

[55] 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. 2024. 1