Drive-R1: Bridging Reasoning and Planning in VLMs for Autonomous Driving with Reinforcement Learning

Yue Li[1*], Meng Tian[2], Dechang Zhu[2], Jiangtong Zhu[2], Zhenyu Lin[3], Zhiwei Xiong[1†], Xinhai Zhao[3†]

[1] University of Science and Technology of China
[2] Yinwang Intelligent Technology Co. Ltd.
[3] Huawei Noah's Ark Lab
yueli65@mail.ustc.edu.cn, zwxiong@ustc.edu.cn, zhaoxinhai1@huawei.com

Abstract

Large vision-language models (VLMs) for autonomous driving (AD) are evolving beyond perception and cognition tasks toward motion planning. However, we identify two critical challenges in this direction: (1) VLMs tend to learn shortcuts by relying heavily on history input information, achieving seemingly strong planning results without genuinely understanding the visual inputs; and (2) the chain-of-thought (COT) reasoning processes are always misaligned with the motion planning outcomes, and how to effectively leverage the complex reasoning capability to enhance planning remains largely underexplored. In this paper, we start from a small-scale domain-specific VLM and propose Drive-R1, designed to bridge the scenario reasoning and motion planning for AD. Drive-R1 first undergoes the supervised finetuning on an elaborate dataset containing both long and short COT data. Drive-R1 is encouraged to reason step-by-step from visual input to final planning decisions. Subsequently, Drive-R1 is trained within a reinforcement learning framework that incentivizes the discovery of reasoning paths that are more informative for planning, guided by rewards based on predicted trajectories and meta actions. Experimental evaluations on the nuScenes and DriveLM-nuScenes benchmarks demonstrate that Drive-R1 achieves superior performance compared to existing state-of-the-art VLMs. We believe that Drive-R1 presents a promising direction for bridging reasoning and planning in AD, offering methodological insights for future research and applications.

Datasets — https://github.com/Depth2World/Drive-R1

Introduction

Autonomous driving (AD) systems aim to enable vehicles to perceive, understand, and interact with their environments in a safe and intelligent manner. Among the core modules in AD pipelines, motion planning plays a central role in determining the future actions, balancing the safety, efficiency, and comfort in real-world driving scenarios. Given observations of the environment and other agents, trajectory prediction directly influence the subsequent low-level control.

Traditional motion planning methods often rely on manually crafted rules (Chen et al. 2015; Fan et al. 2018) that operate under simplified assumptions of the environment and agent behaviors. While these approaches offer interpretability and robustness in structured scenarios, they typically struggle to handle uncertainty, multi-agent interaction, and diverse traffic patterns. Recently, deep learning-based methods (Hu et al. 2022, 2023; Jiang et al. 2023) have shown remarkable success in trajectory prediction by leveraging large-scale driving datasets. These methods, comprised of encoder-decoder architectures or spatio-temporal transformers, model the complex agent dynamics and social interactions. The trajectory prediction lacks the interpretability and still faces limitations in reasoning under ambiguous contexts, adapting to open-world conditions and long-tailed events. The emergence of large vision-language models (VLMs) have introduced new opportunities for enhancing AD systems. Recent methods (Xu et al. 2024; Nie et al. 2024; Marcu et al. 2024; Mao et al. 2023; Tian et al. 2024; Sima et al. 2024) have demonstrated promising results in scene perception, description, and decision-making with analysis in open-formed visual question answer tasks. Further, the methods (Mao et al. 2023; Huang et al. 2024; Wang et al. 2024a; Jiang et al. 2025) extend the perception and cognition tasks to motion planning tasks, with some output interpretable decision processes.

However, several fundamental limitations remain insufficiently addressed in current VLM-based planning systems. *1) The utilization of visual-grounded response in motion planning is limited or even entirely absent.* Recent VLM-based approaches (Tian et al. 2024; Hwang et al. 2024) achieve strong open-loop metrics by predicting trajectories from image-text inputs, often with short or no chain-of-thought (COT) reasoning. Early GPT-driver (Mao et al. 2023) revealed that transforming all perceptual and historical information into textual inputs and using a pure large language model alone can already produce competitive planning performance. To further probe this, we train a general VLM to predict trajectories without COT supervision. At the test time, we ablate the visual input entirely and find that

[*] The work was done during Yue Li's internship at Huawei Noah's Ark Lab.
[†] Corresponding authors.

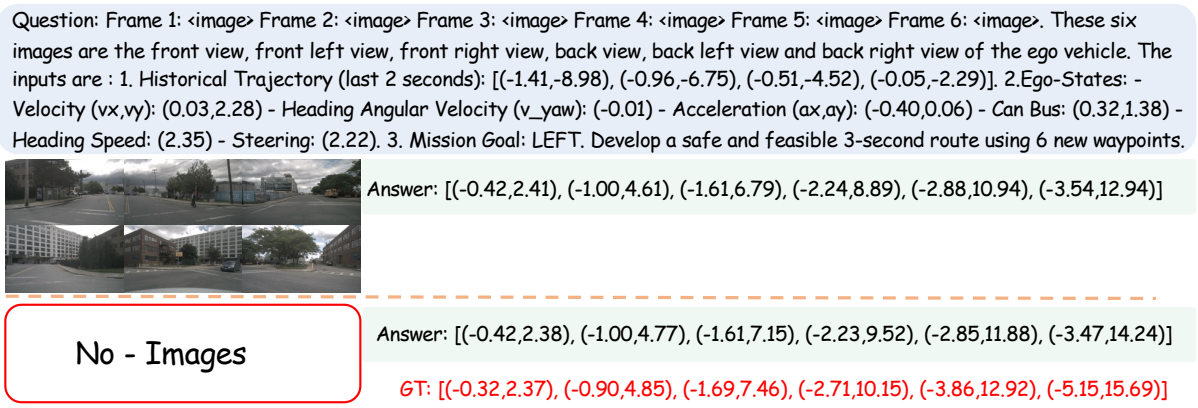


Figure 1: Inference results with and without visual inputs from the model which is trained to predict trajectory without chain of thoughts.

the model performs comparably or even better than with full multi-model inputs. The observation indicates that VLMs for motion planning under-utilize the visual modality and heavily rely on textual priors, especially historical motion cues, raising concerns about their visual grounding and generalization. *2) The COT reasoning traces are always misaligned with the motion planning outcomes.* Leveraging the reasoning capability to enhance planning performance remains largely underexplored. Recent methods (Tian et al. 2024; Huang et al. 2024) engages in sequential question-answering to arrive at the final trajectory prediction. While such methods introduce interpretable intermediate steps, the reasoning remains loosely coupled with planning. We further observe that when a domain-specific (DS) VLM is trained on motion planning datasets with CoT reasoning, it often falls into a reasoning trap. First, the reasoning patterns learned from CoT data especially designed for complex scenarios may introduce unnecessary analysis in simple cases, leading to overthinking and ultimately injecting noise into the planning output. Second, even manually annotated CoT cannot guarantee precise alignment with the ground-truth trajectories, as natural language reasoning tends to be coarse-grained and ambiguous compared to the fine-grained numerical representation required for planning (Jiang et al. 2025).

To address the aforementioned challenges and bridge the gap between scenario reasoning and trajectory-level motion planning in AD, we introduce Drive-R1 tailored for vision-language reasoning and trajectory prediction. We begin with a general VLM, InternVL2 (Chen et al. 2024b), and adapt it to the AD domain by post-training on a large-scale, self-collected dataset comprising 3 million samples. This DS model is endowed with strong perception and scene understanding capabilities, forming a foundation for downstream planning tasks.

To enable reasoning-aware planning, we construct a structured annotation pipeline that generates CoT data according to key domains in real-world AD (Li et al. 2025a), including traffic knowledge understanding, general element recognition, traffic graph generation, target attribute comprehension, and ego decision-making and planning. The resulting CoT dataset contains approximately 4,000 samples, categorized into short and long CoT based on the complexity of the driving scenarios: short CoT corresponds to relatively simple situations that require minimal deliberation, whereas long CoT is designed for complex, multi-agent, or rule-intensive scenes demanding richer step-by-step reasoning. During the supervised learning stage, Drive-R1 is trained on the elaborate dataset to learn to reason from visual observations toward final planning outputs in an interpretable and structured manner. This stage is crucial for encouraging grounded reasoning and mitigating the tendency to overfit to historical trajectory patterns or exploit dataset shortcuts.

To further align the textual reasoning and numerical trajectory planning, we introduce the reinforcement learning (RL) inspired by the success of recent RL approaches (Guo et al. 2025; Huang et al. 2025). Specifically, Drive-R1 employs the Group Relative Policy Optimization (GRPO), which performs optimization over a set of candidate solutions. The relative optimization mechanism is particularly suitable for motion planning, where multiple plausible trajectories may exist under the same driving scenario. By leveraging comparisons across diverse candidates, GRPO encourages the model to discover reasoning paths that generalize well across variations, rather than overfitting to a single deterministic trajectory, thereby enhancing both planning robustness and generalization. The reward design in GRPO integrates four components: trajectory accuracy, meta-action correctness, repetition penalty, and output format compliance. Among them, the trajectory reward captures outcome-level planning quality, while the meta-action reward reflects the reasoning process quality. These two reward signals are complementary, further promoting effective alignment between reasoning and planning within the Drive-R1.

We conduct extensive experiments on both the nuScenes (Caesar et al. 2020) dataset and the DriveLM-nuScenes (Sima et al. 2024) dataset. Our proposed Drive-R1 achieves state-of-the-art performance on the trajectory prediction task, demonstrating its effectiveness in visual-grounded motion planning. Furthermore, we perform

comprehensive ablation studies on DriveLM-nuScenes, investigating the impact of various components, including the GRPO for models under different phases, the number of rollouts, and the influence of different reward functions. Our contribution can be summarized as follows:

- We identify two key challenges in applying VLMs to motion planning: (i) the over-reliance on historical textual inputs leads to shortcut learning, weakening the visual grounding; and (ii) the misalignment between reasoning chains and planning outputs hinders effective integration of interpretability and decision quality.
- We propose Drive-R1, a DS VLM tailored for AD, which connects visual-grounded reasoning to trajectory planning. Our approach incorporates supervised learning on a carefully constructed dataset containing both long and short CoT annotations, followed by RL to further align reasoning quality with planning performance.
- Extensive experiments on nuScenes and DriveLM-nuScenes demonstrate that Drive-R1 achieves state-of-the-art results on trajectory prediction.

While our work represents a straightforward exploration of integrating VLM into the motion planning, the insights gained from Drive-R1 may offer valuable guidance for future efforts toward the practical deployment of AD VLMs.

Related Work

Vision-language Models for Autonomous Driving

The integration of VLMs into AD has recently gained significant attention, aiming to unify perception, reasoning, and planning within a single framework. Existing works in this field can be broadly divided into two categories: scene reasoning-oriented models, and planning and control-oriented models. The first focuses on scene understanding and reasoning (Marcu et al. 2024; Ma et al. 2024; Sima et al. 2024; Nie et al. 2024; Ding et al. 2024), where VLMs are used to analyze visual environments through natural language, often leveraging question-answering or chain-of-thought reasoning to enhance transparency and trustworthiness. Planning and control-oriented models (Tian et al. 2024; Wang et al. 2024a; Xu et al. 2024; Pan et al. 2024; Chen et al. 2024a; Shao et al. 2024), on the other hand, aim to directly generate actionable outputs such as trajectories or control signals from visual and linguistic inputs. These systems often leverage large-scale data and unified modeling to perform planning implicitly within the language model, with or without intermediate reasoning steps. In this paper, we focus on trajectory prediction and find that models can achieve competitive planning performance even with limited or no visual input, suggesting a potential over-reliance on linguistic or historical features and insufficient grounding in visual observations.

Reinforcement Learning

RL has played a pivotal role in the recent evolution of large language models, particularly in aligning model outputs with human preferences or task-specific objectives. Early developments, such as proximal policy optimization (Schulman et al. 2017) and direct policy optimization (Rafailov et al. 2023) have been widely adopted in general-purpose LLMs to improve response helpfulness and safety from human feedback, demonstrating that large models could benefit from post-training optimization beyond supervised learning, enabling them to reason and act in more aligned and consistent ways. Recent Group Relative Policy Optimization (GRPO) (Guo et al. 2025) proposes a group-wise relative optimization strategy, which compares the relative merits of multiple output candidates instead of optimizing based solely on absolute reward values. GRPO has shown strong potential in complex reasoning tasks by encouraging models to explore interpretable thought processes rather than shortcutting to answers. Building on this, RL has been extended to VLMs to enhance their abilities to perform multi-step reasoning grounded in visual inputs (Huang et al. 2025; Jiang et al. 2025; Lai et al. 2025). AlhpaDrive (Jiang et al. 2025) introduced RL to high-level planning and ReCogDrive (Li et al. 2025b) claimed the gap between the discrete language space and the continuous action space. Aligning textual reasoning with numerical outputs like trajectories in AD presents unique challenges, requiring designs that balance process-level and result-level precision. In this paper, our work still pursues the alignment between reasoning and planning in the discrete space.

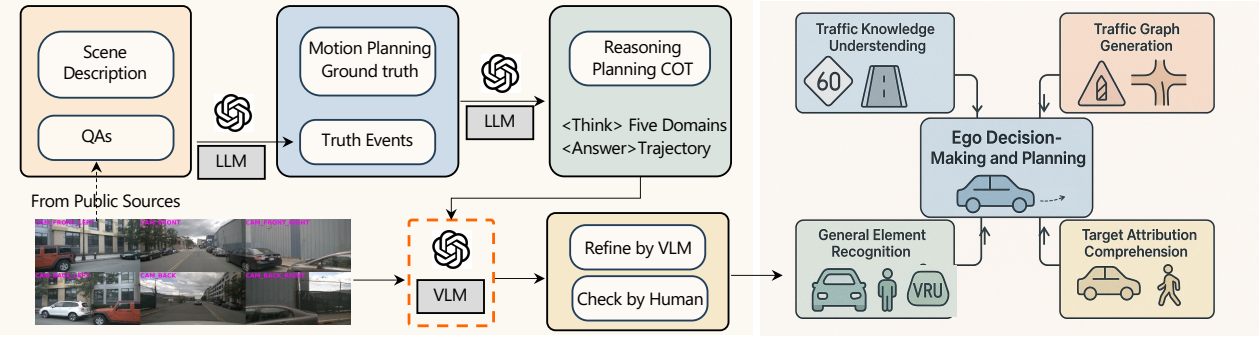
Method

Drive-R1 aims to bridge scenario-level reasoning and trajectory planning for AD through a combination of SFT and RL. We begin by introducing the construction of the Reasoning-Planning chain of thought (RP-CoT) dataset, which encodes intermediate reasoning steps aligned with planning outcomes. Then we detail the supervised training phase, highlighting the initial capabilities the model must acquire to address the challenges discussed above. Finally, we describe the RL procedure, which leverages carefully designed reward functions to further align textual reasoning with numerical trajectory prediction, enhancing both interpretability and planning performance.

RP-COT Data Annotation

Following the five key domains identified in (Li et al. 2025a) as fundamental abilities in motion planning, i.e., traffic knowledge understanding, general element recognition, traffic graph generation, target attribute comprehension, and ego decision-making and planning, we construct the RP-CoT dataset. RP-CoT is designed to bridge high-level reasoning and low-level trajectory prediction. The scenes are selected from nuScenes (Caesar et al. 2020). Each annotation sample in RP-CoT includes step-by-step textual reasoning that reflects a structured understanding of the driving scene, ultimately grounded in a precise trajectory decision.

As shown in Fig.2, the annotation pipeline is semi-automatic. We begin by collecting driving scenes from publicly available sources (Sima et al. 2024; Qian et al. 2024; Inoue et al. 2024), which are annotated with scene descriptions and question-answer (QA) pairs. According to the hundreds of QAs of the scene, the language model ChatGPT first



(a) Annotation Pipeline (b) Reasoning-Planning Chain of Thoughts

Figure 2: The RP-COT data annotation pipeline.

generates the truth events, which are structured representations of the underlying reasoning rationale. Next, based on the ground-truth events and the motion planning information (trajectory, ego status, meta action), ChatGPT generates RP-COT data through the five domains. Each sample includes the `<think></think>` section that explains reasoning steps and the `<trajectory></trajectory>` section that specifies the future trajectory (6 points within 3 seconds). To ensure the generated RP-CoTs are grounded in visual reality, the VLM, GPT-4o, is employed to refine these outputs by aligning them with scene content. Finally, all annotations are checked by human annotators to guarantee consistency, correctness, and planning validity.

Our annotation pipeline systematically decomposes the visual-linguistic information into reasoning stages aligned with the aforementioned domains. This structured format enables the model to learn interpretable reasoning paths that progressively lead to planning actions, laying a strong foundation for subsequent learning stages.

Supervised Fine-tuning Phase

As discussed above, the utilization of visual-grounded response in motion planning is limited or even absent. As shown in Table 1, directly training a general VLM (Internvl2-4B) to output trajectories without CoT supervision can yield surprisingly competitive performance. However, we observe a counterintuitive outcome: the model performs better when visual inputs are ignored, indicating a strong reliance on historical textual context over visual perception. We attribute this phenomenon to two key factors: (1) the model lacks sufficient familiarity with DS tasks in AD and (2) the model are more sensitive to historical motion cues than to scene-level visual information. To address this, we first perform full-parameter finetuning of an InternVL2-4B model on a large-scale, DS dataset comprising 3 million AD QAs, which are collected from public sources (Parikh et al. 2024; Lu et al. 2025; Marcu et al. 2024; Sima et al. 2024; Cao et al. 2024; Li et al. 2022; Malla et al. 2023; Kim et al. 2019; Ma et al. 2024; Wang et al. 2023; Guo et al. 2023; Xu et al. 2024; Mao et al. 2023). The DS model from the first SFT stage significantly mitigates the overreliance on historical information and enhances its general understanding of AD scenarios. Nevertheless, the gap between visual-informed and vision-agnostic reasoning remains narrow. Further, we incorporate the previously constructed RP-CoT dataset into the second SFT stage. Through supervised CoT supervision, the model is encouraged to form visual-grounded reasoning paths across key domains, solving the dependency on textual history information and thereby promoting more robust, perception-aware planning behavior.

On the other hand, the COT reasoning traces are always misaligned with the motion planning outcomes. The experimental results in Table 1 show that applying long CoT supervision during the SFT stage lead to a decline in performance compared to directly supervising the final trajectory output. Interestingly, similar observations occur in other domains. Recent researches (Wang et al. 2024b; Tan et al. 2025) report that for tasks involving spatial reasoning or numerical sensitivity, models trained with CoT supervision often underperform compared to those trained with direct answer supervision. We hypothesize that the observed performance degradation may stem from two primary factors: (1) the limited representation capacity of small-scale models, which restricts their ability to accurately encode and utilize complex reasoning paths (Li et al. 2025c) and (2) the differing tolerance to errors of the models between textual and numerical outputs. Specifically, reasoning texts generated during CoT supervision may contain semantic inconsistencies or hallucinations, either due to imperfect annotation quality or intrinsic limitations of the model. While such errors may have negligible impact on the interpretability or plausibility of the textual reasoning itself, they can propagate to the numerical prediction stage, e.g., trajectory prediction, where small deviations are amplified into significant planning errors.

To mitigate the negative impact of indiscriminate CoT supervision, we introduce a fast-and-slow thinking strategy in the second SFT stage. The core idea is to adapt the complexity of reasoning supervision to the difficulty of each driving scenario. Specifically, we categorize CoT supervision into short CoT and long CoT, depending on the reasoning demand: short CoT corresponds to relatively simple scenarios requiring minimal deliberation, while long CoT is designed for complex, multi-agent, or rule-intensive scenes that demand richer step-by-step reasoning. We begin by training a

Models	RP-COT		Trainning Phase		L2(m)↓				Collision↓			
	Long	Short	SFT	RFT	1s	2s	3s	Avg	1s	2s	3s	Avg
BA	×	✓	✓	×	0.25	0.55	0.97	0.59	0.00	0.09	0.56	0.22
BA-WI	×	✓	✓	×	0.25	0.53	0.90	0.56	0.00	0.03	0.48	0.18
DS	×	✓	✓	×	0.18	0.42	0.76	0.45	0.00	0.03	0.46	0.16
DS	✓	×	✓	×	0.24	0.53	0.91	0.56	0.00	0.19	0.61	0.27
DS	✓	✓	✓	×	0.19	0.39	0.67	0.41	0.00	0.03	0.29	0.11
BA	-	-	×	✓	0.37	0.75	1.22	0.78	0.00	0.16	0.84	0.33
DS	-	-	×	✓	0.26	0.55	0.93	0.58	0.00	0.19	0.61	0.27
DS	✓	✓	✓	✓	0.17	0.35	0.60	0.37	0.00	0.00	0.30	0.10

Table 1: The preliminary experimental results validated on 799 samples from DriveLM-nuScenes (Sima et al. 2024). BA and DS are the base and domain-specific VLM models. WI denotes inference without visual inputs.

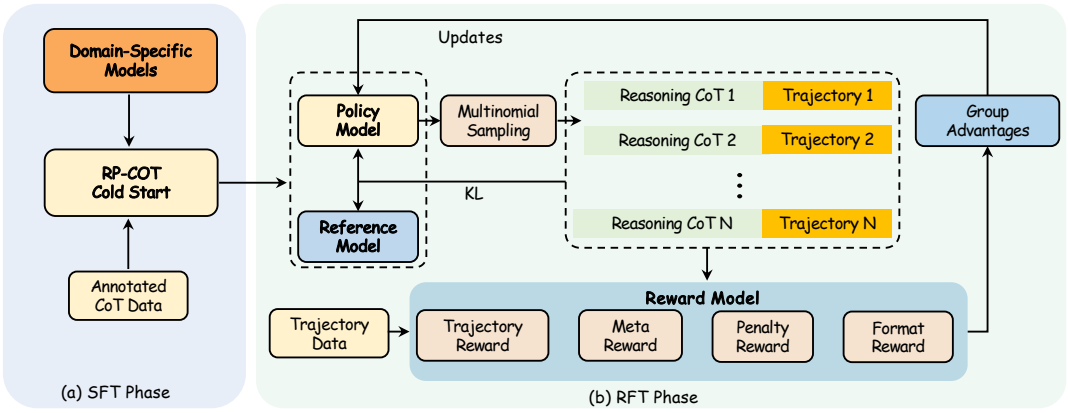


Figure 3: The overview of the proposed Drive-R1, which comprises the supervised fine-tuning (SFT) and reinforcement fine-tuning (RFT) phases.

model to directly generate trajectory outputs without CoT supervision. This specific model is then used to assess the reasoning complexity of each scene, serving as a proxy for scenario difficulty. Scenes with low planning performance are assigned short CoT supervision, while those with high planning performance are paired with long CoT annotations. As shown in Table 1, models fine-tuned with this adaptive fast-and-slow thinking strategy achieve the best overall performance, validating its effectiveness in balancing long and short COT.

Reinforcement Learning Phase

DeepSeek-R1 (Guo et al. 2025) demonstrates that RL frameworks like GRPO can effectively elicit long CoT reasoning abilities of large language models. However, subsequent studies (Yue et al. 2025; Chu et al. 2025) have shown that the reasoning paths produced by RL-finetuned models already exist with high probability in the output distribution of base model, i.e., problems solvable by the RL model can also be addressed by the base model through sufficient sampling. Building upon these insights, we adopt GRPO not as a means to unlock fundamentally new capabilities, but rather as a post-training alignment mechanism to improve the efficiency and consistency for further aligning the reasoning and planning.

Algorithm Specifically, for each question q, GRPO (Guo et al. 2025) samples a group of candidate outputs $\{o_1, o_2, \cdots, o_G\}$ from the old policy $\pi_{\theta_{\text{old}}}$, and subsequently updates the current policy π_θ by maximizing the following objective:

$$\mathcal{J}_{\text{GRPO}}(\theta) = \mathbb{E}_{q,\{o_i\}_{i=1}^{G}\sim\pi_{\theta_{\text{old}}}(O|q)} \left[\frac{1}{G}\sum_{i=1}^{G} w_i A_i - \beta \mathbb{D}_{\text{KL}}(\pi_\theta \| \pi_{\text{ref}}) \right], \tag{1}$$

$$w_i = \min\left(\frac{\pi_\theta(o_i|q)}{\pi_{\theta_{\text{old}}}(o_i|q)}, \text{ clip}\left(\frac{\pi_\theta(o_i|q)}{\pi_{\theta_{\text{old}}}(o_i|q)}, 1-\epsilon, 1+\epsilon \right) \right), \tag{2}$$

$$A_i = \frac{r_i - \text{mean}(\{r_1, r_2, \ldots, r_G\})}{\text{std}(\{r_1, r_2, \ldots, r_G\})}, \tag{3}$$

where the KL loss is calculated by $\mathbb{D}_{\text{KL}}(\pi_\theta \| \pi_{\text{ref}}) = \frac{\pi_{\text{ref}}(o_i|q)}{\pi_\theta(o_i|q)} - \log \frac{\pi_{\text{ref}}(o_i|q)}{\pi_\theta(o_i|q)} - 1$, $\{r_1, r_2, \cdots, r_G\}$ are the rewards of the candidate outputs.

Rewards To better align the reasoning steps with motion planning outcomes, we design a composite reward function that balances both process-level and outcome-level results. The total rewards comprise the following components:

Trajectory Reward measures the accuracy of the predicted trajectory by computing the L2 distance between the predicted trajectory $\hat{\tau}$ and the ground truth τ. The raw distance is then mapped using a sigmoid-based transformation: $R_{\text{traj}} = \frac{2e^{-d}}{1+e^{-d}}$, $d = \|\hat{\tau} - \tau\|_2$.

Meta-Action Reward assesses the high-level planning decisions in the reasoning section, including the short-term lateral and longitudinal decisions. Each contributing 0.5 to the total reward score.

Repetition Penalty penalizes the generation of redundant or repetitive reasoning steps within the CoT to encourage concise and efficient planning rationale (Yeo et al. 2025).

Format Reward ensures correctness of the output structure.

Training Through extensive experiments, we observe that effective RL in the context of motion planning is highly dependent on the model's prior alignment with the AD domain. When applied to models without sufficient domain adaptation, reinforcement signals often result in unstable updates or limited policy improvement, suggesting that the capacity to interpret structured driving scenarios is a prerequisite for successful policy refinement. Consequently, we perform RL on a model that has been supervised via two-stage finetuning, as introduced in the SFT phase. Building on such warm-up, RL further amplifies the synergy between visual-grounded reasoning and motion planning, leading to performance gain observed in our experiments.

Experiments

Datasets and Baselines

In the first SFT phase, the domain-specific data are collected from a diverse set of AD datasets, including (Parikh et al. 2024; Lu et al. 2025; Marcu et al. 2024; Sima et al. 2024; Cao et al. 2024; Li et al. 2022; Malla et al. 2023; Kim et al. 2019; Ma et al. 2024; Wang et al. 2023; Guo et al. 2023; Xu et al. 2024; Mao et al. 2023) with 3 million samples. The QAs are built following the five key domains, and include the single-view, multi-view, and sequential image inputs. In the second SFT phase, RP-COT dataset are construct from the annotations in (Sima et al. 2024; Qian et al. 2024; Inoue et al. 2024) with the number of samples 4,072. The terms short and long are relative: in practice, short samples refer to those without COT annotations, while long samples include CoT reasoning data. When compared on the 6019 validation samples on nuScenes (Caesar et al. 2020), the numbers of short and long RP-COT are 24058 and 4072. When compared on the 799 validation samples on DriveLM-nuScenes (Sima et al. 2024), the numbers of short and long RP-COT are 2036 and 2036. In the RL phase, the samples are selected from those in 4072 RP-COT datasets.

We benchmark Drive-R1 against both end-to-end and vision-language planning baselines. The former includes ST-P3 (Hu et al. 2022), UniAD (Hu et al. 2023) and their modified versions augmented with ego-status inputs (Jiang et al. 2023). The latter set of baselines includes DriveVLM (Tian et al. 2024), RDA-Driver (Huang et al. 2024), OmniDrive (Wang et al. 2024a), and EMMA (Hwang et al. 2024). Notably, prior approaches typically output direct trajectory predictions either without reasoning or with short CoT supervision. In contrast, Drive-R1 produces both reasoning chains and trajectory outcomes in a unified manner, enabling interpretable and context-sensitive planning.

Implementation and Metrics

The SFT training is conducted based on the official codebase of InternVL2 (Chen et al. 2024b). The first-stage SFT is trained on 32 V100 nodes with a batch size of 256, while the second-stage SFT is trained on 16 V100 nodes with a batch size of 128. The RL phase is implemented using the ms-swift framework (Zhao et al. 2025) and trained on 2 V100 nodes with a batchsize of 16 and a rollout of 6. The context length is set to 4096. For evaluation, we adopt the L2 distance and collision rate metrics, following ST-P3 (Hu et al. 2022).

Results

Table 2 presents a comprehensive comparison between Drive-R1 and existing representative baselines. Drive-R1 achieves the lowest average L2 error of 0.31. Although the improvement in L2 distance is relatively modest, it is noteworthy that Drive-R1 is built upon a non-SOTA base model and is capable of generating complete CoT reasoning traces for each planning decision, providing enhanced interpretability and transparency in safety-critical scenarios. In contrast, several end-to-end methods demonstrate competitive L2 metrics, yet suffer from relatively higher collision rates. This suggests that while these models may fit the trajectory well numerically, they may lack robustness in safety-critical aspects of planning. Among VLM-based baselines, Drive-R1 consistently achieves better planning quality and safety. Notably, compared with RDA-Driver (Huang et al. 2024) and OmniDrive (Wang et al. 2024a), our model demonstrates both improved trajectory precision and reduced collision risks, validating the effectiveness of reasoning-aligned trajectory generation.

Ablation Studies

We conduct extensive ablation experiments on the DriveLM-nuScenes (Sima et al. 2024) to investigate the effects of RP-COT input types and RL configurations.

Effect of COT Length in SFT Stage. As shown in Table 1, we evaluate the influence of long and short RP-COTs in the second SFT stage. Models trained with only short RP-COTs or long RP-COTs underperform those trained with both long and short RP-COTs, suggesting that applying a uniform CoT strategy across diverse scenarios is suboptimal. Instead, combining both short and long CoTs better equips the model to handle a wider variety of AD contexts, leveraging both concise and elaborate reasoning chains.

Effectiveness of RL on Different Model Bases. We further assess how RL impacts different model variants, as results summarized in Table 1. The DS model, pre-trained on the first SFT stage, shows a greater performance boost from RL compared to the base model. Moreover, when incorporating long and short RP-COT pattern, the model achieves substantial gains during the RL phase, with noticeable reductions

Models	L2(m)↓				Collision↓			
	1s	2s	3s	Avg	1s	2s	3s	Avg
nuScenes								
ST-P3 (Hu et al. 2022)	1.33	2.11	2.90	2.11	0.23	0.62	1.27	0.71
UniAD (Hu et al. 2023)	0.48	0.96	1.65	1.03	0.05	0.17	0.71	0.31
UniAD-E (Hu et al. 2023)	0.20	0.42	0.75	0.46	0.02	0.25	0.84	0.37
VAD-E (Jiang et al. 2023)	0.17	0.34	0.60	0.37	0.07	0.10	0.24	0.14
DriveVLM (Tian et al. 2024)	0.18	0.34	0.68	0.40	0.10	0.22	0.45	0.27
RDA-Driver (Huang et al. 2024)	0.17	0.37	0.69	0.40	0.01	**0.05**	0.26	0.10
OmniDrive (Wang et al. 2024a)	**0.14**	0.29	0.55	0.33	**0.00**	0.13	0.78	0.30
EMMA (Hwang et al. 2024)	**0.14**	0.29	0.54	0.32	-	-	-	-
Drive-R1 (Ours)	**0.14**	**0.28**	**0.50**	**0.31**	0.02	0.06	**0.19**	**0.09**
nuScenes-DriveLM								
ST-P3 (Hu et al. 2022)	1.28	2.03	2.81	2.04	0.14	0.72	1.28	0.71
GPT-Driver (Mao et al. 2023)	0.22	0.43	0.73	0.46	**0.00**	0.13	0.46	0.19
RDA-Driver (Huang et al. 2024)	0.18	0.38	0.68	0.41	**0.00**	0.06	0.36	0.14
Drive-R1 (Ours)	**0.17**	**0.35**	**0.60**	**0.37**	**0.00**	**0.00**	**0.30**	**0.10**

Table 2: Overall comparison with baselines on the validation dataset.

Rewards			Rollouts	Collision↓			
T.&F.	R.	M.	Nums	1s	2s	3s	Avg
✓	×	×	6	0.06	0.06	0.42	0.18
✓	✓	×	6	0.06	0.06	0.30	0.14
✓	✓	✓	6	0.00	0.00	0.30	0.10
✓	✓	✓	12	0.00	0.03	0.25	0.11
✓	✓	✓	24	0.00	0.03	0.21	0.08

Table 3: Ablation studies of the reward designs and the number of rollouts in GRPO. T., F., R., M. represent trajectory reward, format reward, repetition penalty, and meta-action reward.

in both trajectory deviation and collision rates. These findings highlight the necessity of prior domain alignment before performing RL fine-tuning.

Impact of Reward Design and Rollout Numbers. In Table 3, we evaluate how various reward components and rollout numbers affect model performance. The inclusion of meta-action rewards and repetition penalties leads to consistent improvements in collision rates, highlighting their effectiveness in guiding safer planning behavior. However, for models with relatively small capacity, simply increasing the number of rollouts does not always yield stable or consistent performance gains. For instance, although the collision rate decreases from 0.11 to 0.10 when the number of rollouts increases from 12 and to 24, we observe that the training becomes unstable. It is worth noting that the reported result at 24-rollout is extracted before the onset of training collapse.

Conclusion

In this work, we present Drive-R1, which bridges the structured chain-of-thought reasoning and the trajectory-level motion planning. To address the insufficient visual grounding and the misalignment between reasoning traces and planning outputs, we construct a domain-specific VLM and augment it with a systematically annotated CoT dataset

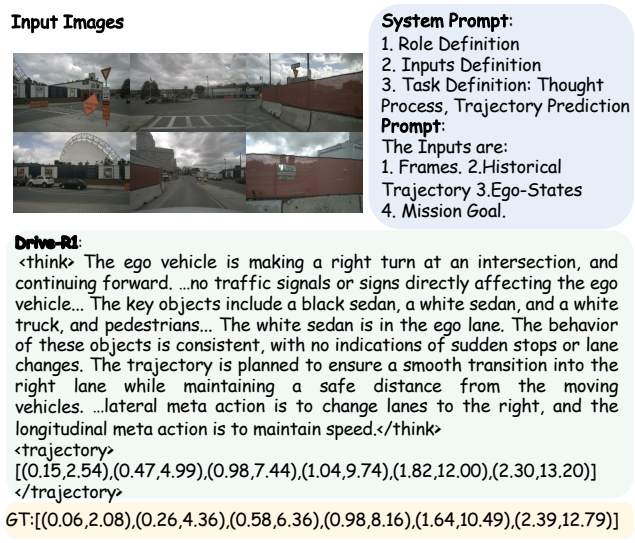


Figure 4: An inference results from our proposed Drive-R1.

spanning five essential reasoning domains. Furthermore, we incorporate a RL phase based on GRPO to optimize planning quality for aligning the reasoning process with trajectory outcomes. Comprehensive experiments validate the effectiveness of our proposed method. Drive-R1 achieves state-of-the-art performance on trajectory prediction tasks while offering interpretable and structured reasoning capabilities. Drive-R1 represents an early exploration toward bridging high-level cognitive reasoning and low-level trajectory planning in AD. In addition, we conduct extensive experiments on large-scale in-house datasets using Ascend 910 hardware platforms, which further verify the generalizability and robustness. We believe that the insights gained from Drive-R1 may offer valuable guidance for future efforts toward the practical deployment of VLMs in AD systems.

Acknowledgments

This work was supported in part by the Fundamental Research Funds for the Central Universities under Grant WK2100000059.

References

Caesar, H.; Bankiti, V.; Lang, A. H.; Vora, S.; Liong, V. E.; Xu, Q.; Krishnan, A.; Pan, Y.; Baldan, G.; and Beijbom, O. 2020. nuscenes: A multimodal dataset for autonomous driving. In *Conference on Computer Vision and Pattern Recognition*.

Cao, X.; Zhou, T.; Ma, Y.; Ye, W.; Cui, C.; Tang, K.; Cao, Z.; Liang, K.; Wang, Z.; Rehg, J. M.; et al. 2024. Maplm: A real-world large-scale vision-language benchmark for map and traffic scene understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 21819–21830.

Chen, C.; Seff, A.; Kornhauser, A.; and Xiao, J. 2015. Deepdriving: Learning affordance for direct perception in autonomous driving. In *Proceedings of the IEEE international conference on computer vision*, 2722–2730.

Chen, Y.; Ding, Z.-h.; Wang, Z.; Wang, Y.; Zhang, L.; and Liu, S. 2024a. Asynchronous large language model enhanced planner for autonomous driving. In *European Conference on Computer Vision*. Springer.

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

Chu, T.; Zhai, Y.; Yang, J.; Tong, S.; Xie, S.; Schuurmans, D.; Le, Q. V.; Levine, S.; and Ma, Y. 2025. Sft memorizes, rl generalizes: A comparative study of foundation model post-training. *arXiv preprint arXiv:2501.17161*.

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

Fan, H.; Zhu, F.; Liu, C.; Zhang, L.; Zhuang, L.; Li, D.; Zhu, W.; Hu, J.; Li, H.; and Kong, Q. 2018. Baidu apollo em motion planner. *arXiv preprint arXiv:1807.08048*.

Guo, D.; Yang, D.; Zhang, H.; Song, J.; Zhang, R.; Xu, R.; Zhu, Q.; Ma, S.; Wang, P.; Bi, X.; et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Guo, Y.; Yin, F.; Li, X.-h.; Yan, X.; Xue, T.; Mei, S.; and Liu, C.-L. 2023. Visual traffic knowledge graph generation from scene images. In *International Conference on Computer Vision*, 21604–21613.

Hu, S.; Chen, L.; Wu, P.; Li, H.; Yan, J.; and Tao, D. 2022. ST-P3: End-to-end Vision-based Autonomous Driving via Spatial-Temporal Feature Learning. In *European Conference on Computer Vision*.

Hu, Y.; Yang, J.; Chen, L.; Li, K.; Sima, C.; Zhu, X.; Chai, S.; Du, S.; Lin, T.; Wang, W.; et al. 2023. Planning-oriented autonomous driving. In *Conference on Computer Vision and Pattern Recognition*.

Huang, W.; Jia, B.; Zhai, Z.; Cao, S.; Ye, Z.; Zhao, F.; Xu, Z.; Hu, Y.; and Lin, S. 2025. Vision-r1: Incentivizing reasoning capability in multimodal large language models. *arXiv preprint arXiv:2503.06749*.

Huang, Z.; Tang, T.; Chen, S.; Lin, S.; Jie, Z.; Ma, L.; Wang, G.; and Liang, X. 2024. Making large language models better planners with reasoning-decision alignment. In *European Conference on Computer Vision*, 73–90. Springer.

Hwang, J.-J.; Xu, R.; Lin, H.; Hung, W.-C.; Ji, J.; Choi, K.; Huang, D.; He, T.; Covington, P.; Sapp, B.; et al. 2024. EMMA: End-to-End Multimodal Model for Autonomous Driving. *arXiv preprint arXiv:2410.23262*.

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

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

Jiang, B.; Chen, S.; Zhang, Q.; Liu, W.; and Wang, X. 2025. Alphadrive: Unleashing the power of vlms in autonomous driving via reinforcement learning and reasoning. *arXiv preprint arXiv:2503.07608*.

Kim, J.; Misu, T.; Chen, Y.-T.; Tawari, A.; and Canny, J. 2019. Grounding Human-To-Vehicle Advice for Self-Driving Vehicles. In *Conference on Computer Vision and Pattern Recognition*.

Lai, Y.; Zhong, J.; Li, M.; Zhao, S.; and Yang, X. 2025. Med-r1: Reinforcement learning for generalizable medical reasoning in vision-language models. *arXiv preprint arXiv:2503.13939*.

Li, K.; Chen, K.; Wang, H.; Hong, L.; Ye, C.; Han, J.; Chen, Y.; Zhang, W.; Xu, C.; Yeung, D.-Y.; et al. 2022. Coda: A real-world road corner case dataset for object detection in autonomous driving. In *European Conference on Computer Vision*. Springer.

Li, Y.; Tian, M.; Lin, Z.; Zhu, J.; Zhu, D.; Liu, H.; Zhang, Y.; Xiong, Z.; and Zhao, X. 2025a. Fine-grained evaluation of large vision-language models in autonomous driving. In *International Conference on Computer Vision*, 9431–9442.

Li, Y.; Xiong, K.; Guo, X.; Li, F.; Yan, S.; Xu, G.; Zhou, L.; Chen, L.; Sun, H.; Wang, B.; et al. 2025b. ReCogDrive: A Reinforced Cognitive Framework for End-to-End Autonomous Driving. *arXiv preprint arXiv:2506.08052*.

Li, Y.; Yue, X.; Xu, Z.; Jiang, F.; Niu, L.; Lin, B. Y.; Ramasubramanian, B.; and Poovendran, R. 2025c. Small models struggle to learn from strong reasoners. *arXiv preprint arXiv:2502.12143*.

Lu, Y.; Yao, Y.; Tu, J.; Shao, J.; Ma, Y.; and Zhu, X. 2025. Can lvlms obtain a driver's license? a benchmark towards reliable agi for autonomous driving. In *AAAI Conference on Artificial Intelligence*, volume 39, 5838–5846.

Ma, Y.; Cao, Y.; Sun, J.; Pavone, M.; and Xiao, C. 2024. Dolphins: Multimodal language model for driving. In *European Conference on Computer Vision*.

Malla, S.; Choi, C.; Dwivedi, I.; Choi, J. H.; and Li, J. 2023. Drama: Joint risk localization and captioning in driving. In *Winter Conference on Applications of Computer Vision*.

Mao, J.; Qian, Y.; Ye, J.; Zhao, H.; and Wang, Y. 2023. Gpt-driver: Learning to drive with gpt. *arXiv preprint arXiv:2310.01415*.

Marcu, A.-M.; Chen, L.; Hünermann, J.; Karnsund, A.; Hanotte, B.; Chidananda, P.; Nair, S.; Badrinarayanan, V.; Kendall, A.; Shotton, J.; et al. 2024. LingoQA: Visual question answering for autonomous driving. In *European Conference on Computer Vision*.

Nie, M.; Peng, R.; Wang, C.; Cai, X.; Han, J.; Xu, H.; and Zhang, L. 2024. Reason2drive: Towards interpretable and chain-based reasoning for autonomous driving. In *European Conference on Computer Vision*. Springer.

Pan, C.; Yaman, B.; Nesti, T.; Mallik, A.; Allievi, A. G.; Velipasalar, S.; and Ren, L. 2024. VLP: Vision Language Planning for Autonomous Driving. In *Conference on Computer Vision and Pattern Recognition*.

Parikh, C.; Saluja, R.; Jawahar, C.; and Sarvadevabhatla, R. K. 2024. IDD-X: A Multi-View Dataset for Ego-relative Important Object Localization and Explanation in Dense and Unstructured Traffic. In *IEEE International Conference on Robotics and Automation*, 14815–14821. IEEE.

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

Rafailov, R.; Sharma, A.; Mitchell, E.; Manning, C. D.; Ermon, S.; and Finn, C. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36: 53728–53741.

Schulman, J.; Wolski, F.; Dhariwal, P.; Radford, A.; and Klimov, O. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Shao, H.; Hu, Y.; Wang, L.; Song, G.; Waslander, S. L.; Liu, Y.; and Li, H. 2024. Lmdrive: Closed-loop end-to-end driving with large language models. In *Conference on Computer Vision and Pattern Recognition*.

Sima, C.; Renz, K.; Chitta, K.; Chen, L.; Zhang, H.; Xie, C.; Luo, P.; Geiger, A.; and Li, H. 2024. DriveLM: Driving with Graph Visual Question Answering. In *European Conference on Computer Vision*.

Tan, H.; Ji, Y.; Hao, X.; Lin, M.; Wang, P.; Wang, Z.; and Zhang, S. 2025. Reason-rft: Reinforcement fine-tuning for visual reasoning. *arXiv preprint arXiv:2503.20752*.

Tian, X.; Gu, J.; Li, B.; Liu, Y.; Wang, Y.; Zhao, Z.; Zhan, K.; Jia, P.; Lang, X.; and Zhao, H. 2024. Drivevlm: The convergence of autonomous driving and large vision-language models. *arXiv preprint arXiv:2402.12289*.

Wang, H.; Li, T.; Li, Y.; Chen, L.; Sima, C.; Liu, Z.; Wang, B.; Jia, P.; Wang, Y.; Jiang, S.; et al. 2023. Openlane-v2: A topology reasoning benchmark for unified 3d hd mapping. *Advances in Neural Information Processing Systems*, 36: 18873–18884.

Wang, S.; Yu, Z.; Jiang, X.; Lan, S.; Shi, M.; Chang, N.; Kautz, J.; Li, Y.; and Alvarez, J. M. 2024a. Omnidrive: A holistic llm-agent framework for autonomous driving with 3d perception, reasoning and planning. *CoRR*.

Wang, W.; Chen, Z.; Wang, W.; Cao, Y.; Liu, Y.; Gao, Z.; Zhu, J.; Zhu, X.; Lu, L.; Qiao, Y.; et al. 2024b. Enhancing the reasoning ability of multimodal large language models via mixed preference optimization. *arXiv preprint arXiv:2411.10442*.

Xu, Z.; Zhang, Y.; Xie, E.; Zhao, Z.; Guo, Y.; Wong, K.-Y. K.; Li, Z.; and Zhao, H. 2024. Drivegpt4: Interpretable end-to-end autonomous driving via large language model. *IEEE Robotics and Automation Letters*.

Yeo, E.; Tong, Y.; Niu, M.; Neubig, G.; and Yue, X. 2025. Demystifying Long Chain-of-Thought Reasoning in LLMs. *arXiv preprint arXiv:2502.03373*.

Yue, Y.; Chen, Z.; Lu, R.; Zhao, A.; Wang, Z.; Song, S.; and Huang, G. 2025. Does reinforcement learning really incentivize reasoning capacity in llms beyond the base model? *arXiv preprint arXiv:2504.13837*.

Zhao, Y.; Huang, J.; Hu, J.; Wang, X.; Mao, Y.; Zhang, D.; Jiang, Z.; Wu, Z.; Ai, B.; Wang, A.; et al. 2025. Swift: a scalable lightweight infrastructure for fine-tuning. In *AAAI Conference on Artificial Intelligence*, volume 39, 29733–29735.