AUTODRIVE-R²: INCENTIVIZING REASONING AND SELF-REFLECTION CAPACITY FOR VLA MODEL IN AUTONOMOUS DRIVING

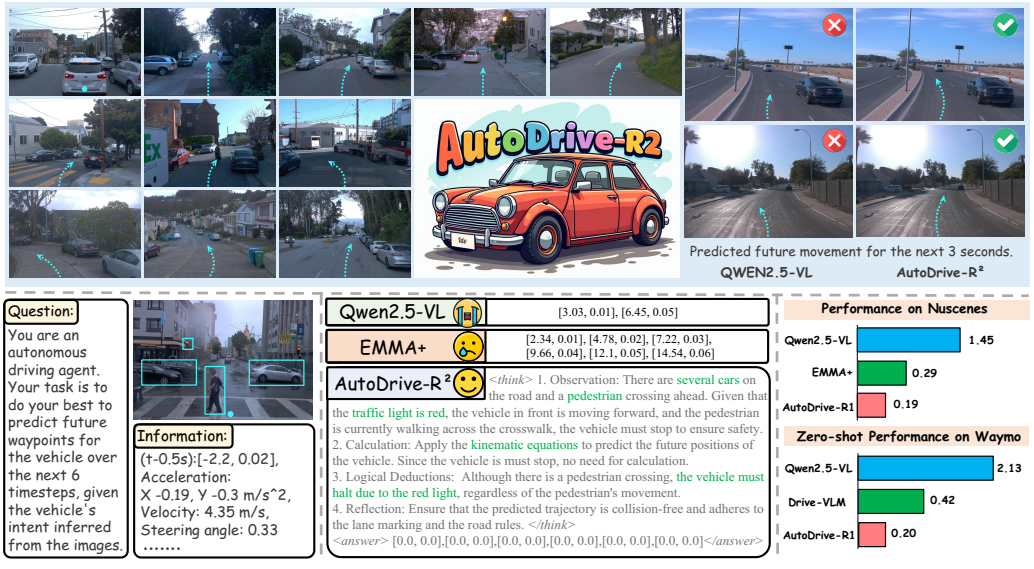
Zhenlong Yuan[1*†], **Chengxuan Qian**[1*], **Jing Tang**[1], **Rui Chen**[1], **Zijian Song**[2], **Lei Sun**[1‡], **Xiangxiang Chu**[1], **Yujun Cai**[3], **Dapeng Zhang**[4§], **Shuo Li**[5]

[1] AMAP, Alibaba Group, [2] Sun Yat-sen University, [3] University of Queensland, [4] Lanzhou University, [5] Case Western Reserve University

[*] Equal contribution [‡] Project Lead [§] Corresponding Author

ABSTRACT

Vision–Language–Action (VLA) models in autonomous driving systems have recently demonstrated transformative potential by integrating multimodal perception with decision-making capabilities. However, the interpretability and coherence of the decision process and the plausibility of action sequences remain largely underexplored. To address these issues, we propose AutoDrive-R², a novel VLA framework that enhances both reasoning and self-reflection capabilities of autonomous driving systems through chain-of-thought (CoT) processing and reinforcement learning (RL). Specifically, we first propose an innovative CoT dataset named nuScenesR²-6K for supervised fine-tuning, which effectively builds cognitive bridges between input information and output trajectories through a four-step logical chain with self-reflection for validation. Moreover, to maximize both reasoning and self-reflection during the RL stage, we further employ the Group Relative Policy Optimization algorithm within a physics-grounded reward framework that incorporates spatial alignment, vehicle dynamic, and temporal smoothness criteria to ensure reliable and realistic trajectory planning. Extensive evaluation results across both nuScenes and Waymo datasets demonstrates the state-of-the-art performance and robust generalization capacity of our method.



[†]Work done during the internship at AMAP, Alibaba Group.

1 INTRODUCTION

Autonomous driving technologies have witnessed rapid advancements in recent years. These systems typically take sensor data as input and then output planning trajectories. Traditional pipelines (Kendall et al., 2019; Chen et al., 2021) usually adopt architectures with separate perception, mapping, prediction, and planning modules. Such design may suffer from two key limitations: error accumulation and lack joint optimization across components, leading to performance degradation. In contrast, modern methods (Hu et al., 2023; Jiang et al., 2023; Chen et al., 2024) unify these complex systems into a single end-to-end paradigm, which naturally offers three main benefits: system simplification, enhanced robustness, and alleviated error accumulation.

However, these end-to-end methods primarily focus on trajectory planning while lacking the contextual reasoning necessary for complex driving scenarios. To address this limitation, recent works integrate Vision-Language Models (VLMs) into autonomous driving systems, leveraging their pre-trained reasoning capabilities to enhance decision-making in challenging situations (Shao et al., 2024a; Wang et al., 2023; Tian et al., 2024). Unlike traditional approaches that train perception-policy modules from scratch, VLM-based methods instead fine-tune pre-trained models by leveraging pre-training on millions of image-text pairs, enhancing vehicles to interpret dynamic traffic situations and develop sophisticated navigation strategies. Despite promising results, current VLM-based systems still struggle with consistently producing accurate planning outputs.

Building upon VLMs, Vision-Language-Action models (VLA) further extend reasoning capabilities to final action prediction, enabling robots and autonomous vehicles to generate precise actions from visual inputs and textual instructions (Yang et al., 2025). This advancement has led to the adoption of similar action generation mechanisms in autonomous driving, with approaches like $\pi 0$ (Black et al., 2024) inspiring the development of action tokenizer that produce precise planning trajectories (Zhou et al., 2025b).

However, current VLA approaches in autonomous driving typically face two critical limitations that hinder their practical deployment: First, existing trajectory generation framework often produce physically infeasible outputs. Existing approaches that directly generate textual commands or waypoints via VLMs frequently produce physically-infeasible outputs and exhibit model collapse. While intermediate representations like meta-actions or latent action tokens have been proposed to mitigate these issues, these designs violate the end-to-end optimization principle and significantly increase model complexity overhead. Second, current systems demonstrate inadequate reasoning capabilities for complex driving scenarios. Since most methods employ simplistic reasoning strategies, they fail to account for both complicated road condition and vehicle kinematic constraints, resulting predicted trajectories significantly deviate from real-world requirements. These limitations underscore the critical need for a novel VLA framework that balances architectural simplicity, robust contextual understanding, and strict physical constraints.

To overcome these challenges, we propose AutoDrive-R², a novel VLA framework that enhances both reasoning quality and physical feasibility through a two-stage training approach. Our key insight is that effective autonomous driving requires structured reasoning processes that can be systematically validated and refined. Specifically, to address the inadequacy of contextual reasoning for complex driving scenarios, we first construct nuScenesR²-6K, a chain-of-thought (CoT) dataset for supervised fine-tuning (SFT). nuScenesR²-6K is the first dataset in autonomous driving that stimulates both reasoning and self-reflection capabilities for VLA models. Unlike prior datasets, nuScenesR²-6K provides not only ground-truth trajectories but also the underlying reasoning and self-reflection steps, ensuring both the correctness and causal plausibility of driving behavior.

Furthermore, to resolve the challenge of physically infeasible trajectory generation, we further develop a physics-grounded reward framework tailored to group relative policy optimization (GRPO) of autonomous driving tasks. By explicitly incorporating spatial alignment, vehicle dynamic and temporal smoothness constraints into consideration, our physics-grounded reward enables reinforcement learning to adapt to diverse driving scenarios and vehicle dynamics while maintaining physical feasibility and motion comfort. Comprehensive experiments on nuScenes and Waymo benchmarks demonstrate that AutoDrive-R² achieves state-of-the-art performance. Our key contributions are:

- We introduce AutoDrive-R², a novel VLA framework that enables semantic reasoning with self-reflection step and trajectory planning from visual information and language instructions.

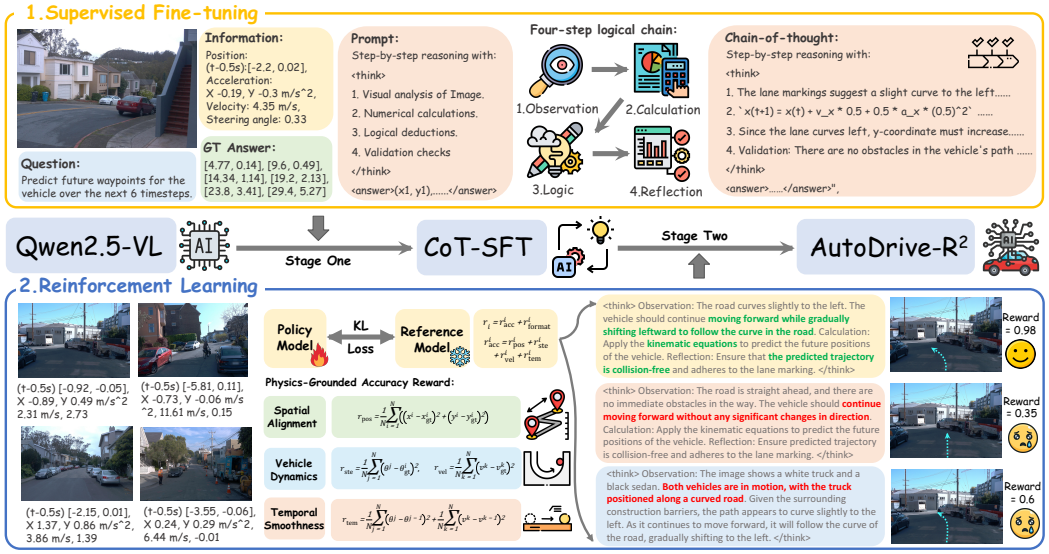


Figure 1: Pipeline of our method. We adopt a two-stage training process. The first stage introduces an innovative CoT dataset named nuScenesR²-6K for SFT. The nuScenesR²-6K adopts a four-step logical chain with self-reflection to generate valuable chain-of-thought data. The second stage proposes an novel physics-grounded reward framework for RL optimization, which incorporates spatial alignment, vehicle dynamic, and temporal smoothness for reliable trajectory planning.

- We propose nuScenesR^2-6K, an innovative CoT dataset adopting a four-step logic chain with self-reflection to help build foundational perception capabilities after SFT.
- We propose a RL-based post-training method based on GRPO, which incorporates physics-grounded rewards as constraint to refine planning trajectory for diverse scenes.

2 METHODOLOGY

Overview. In this section, we present an overview of AutoDrive-R². The target of trajectory planning task requires the model to forecast a vehicle's future motion based on its historical sensor data and contextual information. Formally, given a sequence of historical vehicle states H (including position, acceleration, velocity, steering angle, etc.) and its camera image F, the model M outputs predicted bird's-eye view (BEV) trajectory coordinates T over the next 3 seconds at 0.5-second intervals, defined as $T = M(H, F)$.

As depicted in Fig. 2, our training process contains two stages. In stage one, we construct a high-quality dataset nuScenesR^2-6K for cold start to build cognitive bridges between input information and output trajectories through a four-step logical chain with self-reflection for validation. In stage two, we employ a physics-based reinforcement learning framework that integrates spatial alignment, vehicle dynamic and temporal smoothness to ensure physically feasible trajectory generation.

2.1 LOGISTIC COT DATASET WITH SELF-REFLECTION

The success of VLA models in autonomous driving critically depends on their ability to produce both interpretable reasoning and physically feasible actions. However, existing training approaches often fail to establish this dual requirement, leading to models that either lack explainable decision-making processes or generate unrealistic trajectories. To investigate this challenge, we initially explored direct reinforcement learning optimization for trajectory planning, following recent advances in reasoning-based RL (Guo et al., 2025). However, preliminary experiments revealed that models trained exclusively on RL exhibited significant degradation in trajectory planning compared to models pre-trained with SFT before RL. Therefore, we constructed a high-quality cold-start dataset named nuScenesR²-6K to cultivate the model's foundational understanding of trajectory planning.

To this end, we constructed nuScenesR²-6K, a dataset of 6,000 image-trajectory pairs enriched with high-fidelity chain-of-thought (CoT) reasoning. The dataset was created through a meticulous "generate-then-validate" pipeline. First, we curated an initial pool of approximately 8,000 image-trajectory pairs from the nuScenes training set. We then leveraged the Qwen2.5-VL-72B model to synthesize initial CoT reasoning sequences for these pairs. Crucially, to ensure data quality, we employed the closed-source Qwen-VL-Max model as an expert validator to score and review each generated chain. Chains containing factual errors or logical inconsistencies were systematically discarded. This rigorous filtering process yielded the final 6,000 high-fidelity samples for SFT.

Moreover, we observe that many existing approaches rely on universal prompts for problem-to-answer reasoning, lacking structured guidance for rational analysis. While this strategy proves effective for simple tasks, it frequently fails when confronted with complex mathematical or logical problems. To address this limitation, our CoT prompt design systematically decomposes trajectory planning into three interdependent reasoning stages:

- Image-Driven Analysis: Establishing foundational scene understanding (e.g., obstacle and lane localization, traffic sign detection) to anchor subsequent reasoning.
- Physics-based Calculation: Leveraging kinematic equations (e.g., angular momentum conservation) to translate abstract observations into quantifiable predictions.
- Contextual Logic Synthesis: Integrating domain-specific knowledge (e.g., intersection traffic rules) to ensure predictions align with real-world driving regulations.

To further enhance robustness and the correctness of answers, we explicitly introduce a self-reflection phase as the fourth step, inspired by mathematical reasoning frameworks that validate conclusions through backward-checking. This allows the model to verify the coherence of its reasoning and correct potential contradictions. Our prompt implements a four-step logic chain:

Observation → Calculation → Logic → Reflection,

which delivers both systematic and error-resilient reasoning. Ultimately, the nuScenesR²-6K dataset is adopted for supervised fine-tuning Qwen2-VL-7B model, thus yielding our stage-1 model.

2.2 Group Relative Policy Optimization (GRPO)

We follow the GRPO algorithm (Jiang et al., 2025) to train the model. Unlike traditional approaches that rely on critic networks to estimate value functions, GRPO introduces a pairwise comparison mechanism among candidate response. This design not only simplifies the architecture but also reduces computational overhead during training. The methodology begins by generating G distinct candidate responses $o = \{o_1, \ldots, o_G\}$ for a given input question q through policy sampling. For our specific task, we implement two rule-based verifiable reward functions to assess response quality:

Accuracy Reward To better adapt to trajectory planning task, we define a physics-grounded accuracy rewards r_{acc} which integrates spatial, kinematic, and temporal constraints for evaluation. Details are specified in the following section.

Format Reward The format reward r_{acc} enforces strict adherence to the required output format. The model must produce responses in the form: *"<think>thinking process here</think><answer>(x_1, y_1), ...,(x_n, y_n)</answer>"*. A value of 1 is assigned if the format is correct, otherwise 0. In summary, the total reward for a response o_i is calculated as: $r_i = r^i_{\text{acc}} + r^i_{\text{format}}$. To quantify the relative quality of all responses $\{r_1, \ldots, r_G\}$, GRPO normalizes these scores by subtracting the group mean and dividing by the standard deviation. Consequently, the advantage for each response can be formulated by:

$$A_i = \frac{r_i - \text{mean}(\{r_i\}_{i=1}^G)}{\text{std}(\{r_i\}_{i=1}^G)}, \tag{1}$$

where A_i is the relative advantage of the i-th answer. Then the optimization objective further incorporates a regularization term to ensure the updated policy π_θ remains close to the original reference

policy π_{ref}. This is achieved by adding a KL-divergence term $D_{\text{KL}}(\cdot \,\|\, \cdot)$ to the loss function:

$$\mathcal{L}_{GRPO}(\theta) = -\mathbb{E}[q\sim P(Q), \{o_i\}_{i=1}^{G}\pi_{\theta old}(O|q)]\frac{1}{G}\sum_{i=1}^{G} \left(min\Big(\frac{\pi_\theta(o_i|q)}{\pi_{\theta old}(o_i|q)} * Adv_i, clip\Big(\frac{\pi_\theta(o_i|q)}{\pi_{\theta old}(o_i|q)}, 1-\epsilon, 1+\epsilon\Big) * Adv_i\Big) - \beta\mathbb{D}_{KL}(\pi_\theta||\pi_{ref})\right), \tag{2}$$

$$\mathbb{D}_{KL}(\pi_\theta||\pi_{ref}) = \frac{\pi_{ref}(o_i|q)}{\pi_\theta(o_i|q)} - log\frac{\pi_{ref}(o_i|q)}{\pi_\theta(o_i|q)} - 1, \tag{3}$$

where β acts as a hyperparameter to trade-off between exploration and stability during optimization.

2.3 Physics-Grounded Accuracy Rewards in GRPO

In autonomous driving, traditional reward function designs often focus solely on trajectory position error, while neglecting the complex constraints in geometric, dynamical, and temporal dimensions. To address this issue, we propose a physics-grounded reward framework that integrates spatial alignment, vehicle dynamics, and temporal continuity to comprehensively guide the model in generating safe, feasible, and comfortable driving strategies. This multi-dimensional approach not only ensures geometric accuracy but also explicitly incorporates the physical limitations of real-world vehicles and the perceptual requirements for motion smoothness, creating a holistic optimization objective.

Spatial Alignment: Balancing Global Maneuverability. The foundation of any trajectory reward function lies in its ability to align predicted paths with target routes. We define a spatial accuracy term r_{pos} as the mean squared Euclidean distance between predicted and ground-truth coordinates:

$$r_{\text{pos}} = \frac{1}{N}\sum_{i=1}^{N}\left((x^i - x^i_{\text{gt}})^2 + (y^i - y^i_{\text{gt}})^2\right), \tag{4}$$

where N denotes the number of time steps, x^i, y^i represent predicted coordinates at the i-th time step, while $x^i_{\text{gt}}, y^i_{\text{gt}}$ correspond to the ground-truth values. This formulation prioritizes global path adherence by penalizing deviations across all time steps, ensuring the vehicle remains on the intended route. However, focusing only on minimizing position error may produce physical-implausible results. For instance, strictly following the shortest path might bring about abrupt steering or acceleration, which not only violate vehicle kinematics but also compromise passenger comfort. To balance geometric precision with practical feasibility, we introduce additional constraints derived from vehicle dynamics.

Vehicle Dynamics: Bridging Perception and Control. Autonomous driving systems must respect the real-world physical limitations, which are governed by steering kinematics and longitudinal dynamics. Ignoring them may result in trajectories that are impossible to execute (e.g., requiring infinite torque for abrupt steering changes) or uncomfortable for passengers. To ensure kinematic feasibility, we penalize deviations in steering angles through the following term r_{ste}:

$$r_{\text{ste}} = \frac{1}{N}\sum_{j=1}^{N}\left(\theta^j - \theta^i_{\text{gt}}\right)^2, \tag{5}$$

where θ^j and θ^j_{gt} respectively denotes the predicted and corresponding ground-truth steering angle at j-th time step. Additionally, we address unphysical acceleration/braking patterns by introducing an additional velocity constraint term:

$$r_{\text{vel}} = \frac{1}{N}\sum_{k=1}^{N}\left(v^k - v^k_{\text{gt}}\right)^2, \tag{6}$$

where v^k and v^j_{gt} respectively represents the predicted and corresponding ground-truth velocity at the k-th time step. In summary, both r_{ste} and r_{vel} enforce compliance with vehicle-specific constraints, ensuring generated trajectories are both physically realizable and socially acceptable in

Table 1: Trajectory L2 errors and collision rates on the nuScenes dataset.

Method	L2 Error (m) ↓				Collision Rate (%) ↓			
	1s	2s	3s	Avg.	1s	2s	3s	Avg.
Open-source Generalist Vision Language Models								
Llama-3.2-11B-Vision (Touvron et al., 2023)	1.54	3.31	3.91	2.92	-	-	-	-
DeepSeek-VL2-16B (Liu et al., 2025)	0.66	1.68	2.92	1.75	-	-	-	-
LLaMA-3.2-11B-Vision (Touvron et al., 2023)	0.52	1.42	2.68	1.54	-	-	-	-
Qwen-2.5-VL-3B (Bai et al., 2025)	2.69	4.16	5.78	4.21	0.32	0.54	0.78	0.55
Qwen-2.5-VL-7B (Bai et al., 2025)	0.46	1.33	2.55	1.45	0.10	0.18	0.24	0.17
Training-based Driving Specialists (Existing Methods)								
UniAD (Hu et al., 2023)	0.42	0.64	0.91	0.66	0.62	0.58	0.63	0.61
VAD (Jiang et al., 2023)	0.17	0.34	0.60	0.37	0.07	0.17	0.41	0.22
BEV-Planner (Li et al., 2024)	0.16	0.32	0.57	0.35	0.00	0.29	0.73	0.34
Ego-MLP (Zhai et al., 2023)	0.15	0.32	0.59	0.35	-	-	-	-
Ours and Key Competitors (Specialized Driving Models)								
DriveVLM (Tian et al., 2024)	0.18	0.34	0.68	0.40	0.10	0.22	0.45	0.27
OmniDrive (Wang et al., 2024)	0.14	0.29	0.55	0.33	0.00	0.13	0.78	0.30
DriveVLM-Dual (Tian et al., 2024)	0.15	0.29	0.48	0.31	0.05	0.08	0.17	0.10
EMMA (Hwang et al., 2024)	0.14	0.29	0.54	0.32	-	-	-	-
EMMA+ (Hwang et al., 2024)	0.13	0.27	0.48	0.29	-	-	-	-
Impromptu-VLA	0.13	0.27	0.53	0.30	-	-	-	-
AutoDrive-R² 3B	0.35	0.49	0.62	0.49	0.13	0.44	0.87	0.48
AutoDrive-R² 7B	0.13	0.19	0.25	0.19	0.00	0.07	0.12	0.07

mixed traffic scenarios. These constraints explicitly bridge the gap between perception-driven planning and actuator-level control, ensuring that predicted trajectories align with the physical boundaries while maintaining ride quality.

Temporal Smoothness: Ensuring Navigation Reliability Temporal discontinuities in trajectory predictions fundamentally undermine the reliability of autonomous driving systems. When steering or acceleration commands exhibit sudden jumps between time steps, the predicted trajectories may lose coherence, which further compromises the system's ability to maintain stable, predictable motion patterns required for safe navigation. To address this, we introduce a temporal smoothness term r_{tem} that penalizes rapid variations in consecutive control signals:

$$r_{\text{tem}} = \frac{1}{N}\sum_{j=1}^{N}\left(\theta^{j} - \theta^{j-1}\right)^2 + \frac{1}{N}\sum_{k=1}^{N}\left(v^{k} - v^{k-1}\right)^2. \tag{7}$$

Such design ensures temporal coherence of predicted trajectories. By explicitly constraining the rate of change in both steering and velocity, the reward function filters out unstable oscillations that could destabilize the vehicle's state estimation. This regularization effect strengthens the model's ability to generalize across diverse driving scenarios while maintaining safety margins during execution.

Integrated Reward Function. The final reward synthesizes all dimensions with learnable weights:

$$r_{\text{acc}} = \lambda_{\text{pos}} \cdot r_{\text{pos}} + \lambda_{\text{ste}} \cdot r_{\text{ste}} + \lambda_{\text{vel}} \cdot r_{\text{vel}} + \lambda_{\text{tem}} \cdot r_{\text{tem}}. \tag{8}$$

Here, $\lambda_{\text{pos}}, \lambda_{\text{ste}}, \lambda_{\text{vel}}, \lambda_{\text{tem}}$ are learnable coefficients that balance trade-offs between competing objectives. We set them all equal one in experiments. This holistic formulation ensures the model generates trajectories that are geometrically accurate, dynamically feasible, and temporally smooth, addressing the multifaceted challenges of autonomous driving.

3 EXPERIMENT

3.1 EXPERIMENTAL SETTINGS

Datasets. For training, we adopt nuScenesR^2-6K dataset, which contains 6k image-trajectory sample pairs, each includes a front-view image and a 3-second trajectory planning with 0.5-second intervals. The Qwen2.5-VL-7B model is fine-tuned on these samples for SFT to establish foundational perception capabilities before RL. For evaluation, our method is tested on nuScenes and Waymo datasets, both offering comprehensive autonomous driving data. The nuScenes dataset contains 1,000 urban driving scenes with six synchronized camera views to support planning tasks. Waymo dataset includes 4,021 driving segments, capturing eight camera views and ego-vehicle trajectories.

Details. We implement experiments on both Qwen2.5-VL-3B and Qwen2.5-VL-7B models. In both stages, the learning rate is set to 5e-7 with an accumulated total batch size of 8. The GRPO is configured with a maximum completion length of 4,096 tokens and samples 6 responses per input. For training, the reinforcement learning (RL) pipeline utilized the TRL framework, and was executed for a total of 750 training iterations and 1 epoch for 18 hours.

Evaluation Metrics. We evaluate performance using both accuracy and safety metrics. For accuracy, we adopt the L2 distance (in meters) between the predicted and ground truth trajectories at 1s, 2s, and 3s, along with the average error. To assess safety, we also report the Collision Rate (%), which measures the frequency of collisions in the predicted paths. For all models, we utilize the official checkpoints and conduct evaluations under the same evaluation codes to ensure fairness. Note that the best and second-best results are highlighted in all tables.

3.2 EVALUATION RESULTS

Results on nuScenes Datasets Table 1 compares the prediction errors among our method and existing approaches on the nuScenes dataset. Notably, our approach consistently achieves the best performance across all time intervals, surpassing current leading methods such as EMMA+, which are trained on substantially larger internal datasets with 103k scenarios. In contrast, our training data consists of only 6k curated CoT samples for stage 1 and another 6k for stage 2, approximately 11.65% the size of EMMA+'s dataset. Furthermore, our model demonstrates significant improvements over Qwen2-VL-7B, reducing L2 errors by 86.9%, despite having less parameter.

Zero-shot performance on Waymo Datasets Moreover, Tab. 2 demonstrates the robust zero-shot capabilities of our model. Specifically, our method respectively reduces L2 errors by 33.3% and 90.7% compared to the latest EMMA+ method and Qwen2-VL-72B baseline models. Overall, our model consistently delivers precise trajectory predictions across multiple datasets, establishing its state-of-the-art performance and generalization capability.

Table 2: Trajectory L2 errors on Waymo.

Method	L2 Error (m) ↓			
	1s	2s	3s	Avg.
Generalist VLMs + Specialized Driving Models				
Qwen-2.5-VL-3B	2.98	5.05	7.38	5.14
Qwen-2.5-VL-7B	1.66	1.82	2.92	2.13
DriveVLM	0.17	0.34	0.75	0.42
EMMA	0.12	0.28	0.61	0.34
EMMA+	0.11	0.25	0.53	0.30
AutoDrive-R² 3B	0.23	0.36	0.51	0.37
AutoDrive-R² 7B	0.11	0.19	0.29	0.20

Model Size In Tab. 1 and Tab. 2, we compare 3B and 7B variants of Qwen2.5-VL within our two-stage training framework to analyze impact of different model size. While the 7B model achieves superior performance with an average L2 error of 0.19m, the 3B version demonstrates a notable improvement than its baseline. The disparity highlights that larger models inherently capture more complex patterns, but the two-stage framework (SFT + GRPO) effectively compensates for the 3B model's limited capacity by enforcing strict trajectory constraints and contextual logic synthesis.

Closed-loop Experiment To evaluate our model's practical decision-making, we conducted closed-loop experiments on the NAVSIM benchmark. As detailed in Table 3, our method sets a new state-of-the-art by outperforming all prior approaches across every metric. Specifically, our model achieves top scores in safety, with a Navigation Completion of 98.3 and Time-to-Collision of 95.6, while also excelling in driving quality by leading in Driving Agent Comfort (94.4) and Ego-Progress (81.6).

Figure 2: Qualitative comparison of trajectory planning performance across Qwen2.5-VL-7B, EMMA+, and our AutoDrive-R² on the nuScenes dataset. Note that blue lines denote predicted trajectories while green lines represent ground truth trajectories.

The most significant advantage is observed in the Planner-Diverse Motion Score (PDMS), where our score of 90.3 surpasses the next-best competitor by over 6 points, showcasing superior human-like planning. These comprehensive results confirm our model's ability to translate its predictive accuracy into robust, safe, and efficient real-world driving actions.

Table 3: Performance on the Closed-loop NAVSIM.

Methods	NC↑	DAC↑	TTC↑	Comfort↑	EP↑	PDMS↑
TransFuser (Chitta et al., 2022)	97.7	92.8	92.8	100	79.2	84.1
UniAD (Hu et al., 2023)	97.8	91.9	92.9	100	78.8	83.4
Para-Drive (Weng et al., 2024b)	97.9	92.4	93.0	99.8	79.3	84.0
AutoDrive-R² 7B	98.5	95.9	95.4	100	82.7	89.1

Visualization In Fig. 2, we present a comparative analysis of our method against other approaches in the nuScenes dataset. Notably, Qwen2.5-VL-7B fails to generate accurate predictions in specific scenarios (e.g., (b) and (d)), whereas EMMA+ exhibits significant trajectory deviation. In contrast, our method consistently achieves more reliable and physically feasible trajectory planning under varying illumination environments and complex motion patterns.

3.3 Ablation Studies

Training Stages Drawing inspiration from DeepSeek-R1-Zero, we first attempt to train the model solely through RL. As shown in Tab. 4, the model purely trained on RL (7B + *RL*) underperforms that of SFT (7B + *SFT*) by 22.2% in average L2 error. We attribute this to model's inability to establish structured reasoning chains, as RL struggles to explore the high-dimensional reasoning space required for multi-step calculations and contextual logic synthesis. This observation validates the necessity of our two-stage training.

Table 4: Ablation studies of trajectory L2 errors on nuScenes dataset for validation.

Method	L2 Error (m) ↓			
	1s	2s	3s	Avg.
Qwen2.5-VL-7B	0.46	1.33	2.55	1.45
Qwen2.5-VL-7B + *SFT*	0.17	0.27	0.36	0.27
Qwen2.5-VL-7B + *RL*	0.21	0.33	0.44	0.33
SFT : w/o. Four.	0.19	0.25	0.32	0.25
SFT : w/o. Self.	0.17	0.23	0.29	0.23
RL : w/o. r_{pos}	0.32	0.53	0.72	0.53
RL : w/o. r_{ste}	0.14	0.20	0.27	0.21
RL : w/o. r_{vel}	0.15	0.21	0.29	0.22
RL : w/o. r_{tem}	0.15	0.23	0.34	0.24
AutoDrive-R² 7B	0.13	0.19	0.25	0.19

Supervised Fine-tuning (*SFT*) In the first stage, the baseline Qwen2.5-VL-7B (7B) achieves an average L2 error of 1.45m, whereas the SFT model (7B + *SFT*) trained on the nuScenesR²-6K dataset reduces this to 0.27m, demonstrating an 81.4% improvement. This significant enhancement highlights the effectiveness of adopting SFT training in establishing foundational reasoning capabilities. Moreover, removing four-step reasoning structure (w/o. Four.) increases the error to 0.25m, indicating a 31.5% degradation compared to AutoDrive-R². Similarly, eliminating self-reflection (w/o. Self.) results in 0.23m error, representing a 21.1% decline relative to AutoDrive-R². This emphasizes the interdependence of both four-step logical chain and self-reflection mechanism in constructing high-quality CoT dataset.

Reinforcement Learning (*RL*) In the second stage, we evaluate the contribution of individual reward components within the physics-grounded framework of AutoDrive-R². Specifically, spatial alignment is critical for maintaining global geometric path accuracy, as its removal (w/o. r_{pos}) increases the error to 0.53m, much higher than the full model. Moreover, steering angle regulation ensures kinematic feasibility by penalizing abrupt changes in steering adjustments, and its absence (w/o. r_{ste}) leads to a 10.5% degradation (0.21m). Additionally, velocity consistency constraints ensure adherence to target speed profiles by penalizing deviations in predicted velocity from ground-truth values, and their exclusion (w/o. r_{vel}) raises the error to 0.22m. Finally, temporal smoothness penalties suppress unstable control patterns by penalizing abrupt changes in steering and velocity across time steps, and their removal (w/o. r_{tem}) results in a 26.3% increase in error (0.24m). By integrating all four components into our physics-grounded reward, AutoDrive-R² achieves an optimal 0.19m L2 error, confirming the necessity of each element in achieving reliable trajectory planning.

Key Hyperparameters In Tab. 5, We conduct a series of ablation studies to analyze the impact of key hyperparameters. First, for the reward weights (λ), we find that uniform weights ($\lambda = (1, 1, 1, 1)$) achieve a lower average L2 error (0.19 m) compared to decaying weights (0.22 m). This indicates that all reward components are equally important for the learning process, leading us to adopt the uniform setting. Next, we analyze the KL-divergence coefficient β. A value of $\beta = 0.04$ achieves the best performance with a 0.19 m error. A smaller value ($\beta = 0.02$) results in a higher error (0.21 m), while a larger value ($\beta = 0.06$) also slightly degrades performance (0.20 m), likely by overly constraining the optimization. Finally, for the number of generations (G), we observe performance improving as G increases from 2 (0.23 m) to 6 (0.19 m). However, increasing G further to 8 yields no additional benefit. Therefore, we set $G = 6$ to balance performance and computation.

Table 5: Ablations on key hyperparameters.

Setting	L2 Error (m) ↓			
	1s	2s	3s	Avg.
Reward Weights (λ)				
$\lambda : (0.4, 0.3, 0.2, 0.1)$	0.15	0.22	0.29	0.22
$\lambda : (1, 1, 1, 1)$ **(Ours)**	0.13	0.19	0.25	0.19
KL-divergence Beta (β)				
$\beta = 0.02$	0.14	0.21	0.27	0.21
$\beta = 0.04$ **(Ours)**	0.13	0.19	0.25	0.19
$\beta = 0.06$	0.13	0.20	0.26	0.20
Number of Generations (G)				
$G = 2$	0.16	0.23	0.31	0.23
$G = 4$	0.14	0.20	0.26	0.20
$G = 6$ **(Ours)**	0.13	0.19	0.25	0.19
$G = 8$	0.13	0.19	0.25	0.19

4 Related Work (Extended Ver. in Appx. B)

Autonomous Driving The evolution of autonomous driving systems has transitioned from modular architectures to end-to-end learning paradigms. Early works like UniAD (Hu et al., 2023) pioneered the integration of sub-tasks into a unified framework. Concurrently, other methods addressed core challenges in prediction and efficiency; for example, Flash (Antonello et al., 2022) focused on creating fast models for real-time application, while DiPA (Knittel et al., 2023) tackled the complex, multi-modal nature of agent interactions. These works highlighted the need for a holistic evaluation beyond simple L2 error, inspiring our new analyses on collision rate and closed-loop performance. Subsequent methods like Para-Drive (Weng et al., 2024a) and BEV-Planner (Li et al., 2024) continued to refine these integrated approaches.

The advent of vision-language models (VLMs) has further transformed the field. DriveVLM (Tian et al., 2024) and DriveMLM (Wang et al., 2023) enabled systems to incorporate linguistic reasoning for decision-making and rationale generation. Alongside performance, explainability has become a critical concern. A systematic review in (Kuznietsov et al., 2024) provides a comprehensive overview of techniques for building trustworthy systems, while works like Interpretable Goal-based Prediction and Planning (Albrecht et al., 2021) explored generating explanations by making the model's goals explicit. In contrast to these post-hoc or explicit-goal methods, our AutoDrive-R² introduces a novel form of *intrinsic* explainability, representing a significant step towards inherently trustworthy AI for autonomous driving.

General VLMs The development of general vision-language models (VLMs) has been driven by the success of large language models (LLMs) in understanding textual data. CLIP (Radford et al., 2021) established a foundational approach by aligning image and text features through contrastive learning, enabling zero-shot generalization. Building on this, BLIP (Li et al., 2022) and its successor BLIP-2 (Li et al., 2023) refined multimodal alignment using contrastive and matching losses to improve contextual grounding. More recent models, such as LLaVA (Liu et al., 2023) and Qwen2.5VL (Bai et al., 2025), have integrated robust LLMs with vision encoders to enhance representation capabilities. The OmniGen2 (Wu et al., 2025) framework further advanced this by introducing dual decoding pathways for text and image generation, while DeepSeek-V3 (Liu et al., 2025) demonstrated efficient inference through a Mixture-of-Experts (MoE) architecture with auxiliary-loss-free load balancing. These advancements highlight the growing synergy between vision and language modalities in multimodal learning.

Reinforcement Learning for Post-Training Reinforcement learning (RL) has emerged as a critical tool for refining model capabilities post-training. Techniques like Proximal Policy Optimization (PPO) (Schulman et al., 2017) have been instrumental in fine-tuning models for complex tasks, as seen in the optimization of GPT (Achiam et al., 2024). Direct Preference Optimization (DPO) (Rafailov et al., 2023) introduces a sampling-free parameterization for reward models, streamlining the fine-tuning process. Reward Fine-Tuning (RFT) (Yuan et al., 2023) has shown particular efficacy in mathematical reasoning, while Guided Reward Policy Optimization (GRPO) (Shao et al., 2024b) enables robust reasoning improvements without external toolkits. The DeepSeek-R1 (Guo et al., 2025) model exemplifies the application of GRPO, achieving state-of-the-art results through physics-informed reward design. These methodologies underscore the importance of RL in aligning model outputs with real-world constraints and user preferences.

5 Conclusion

In this work, we propose AutoDrive-R², a novel VLA framework designed for reasoning-guided trajectory planning in autonomous driving. AutoDrive-R² effectively balances semantic understanding with real-world constraints through a two-stage training framework: (1) a SFT stage adopting the nuScenesR²-6K dataset, which employs a four-step CoT process to cultivate structured reasoning and self-reflection for validation, and (2) a RL stage leveraging GRPO training to refine trajectory planning under physics-grounded rewards. Experiments validate the effectiveness of AutoDrive-R², achieving SOTA performance on nuScenes (34.5% reduction vs. EMMA+) and Waymo (90.7% reduction vs. Qwen2.5-VL-7B), demonstrating strong zero-shot generalization. Future efforts will focus on multi-agent coordination and real-time sensor fusion integration to further improve adaptability in complicated environments.

ETHICS STATEMENT

Our research builds upon publicly available autonomous driving datasets, including nuScenes and Waymo, to construct the nuScenesR²-6K dataset. The manual annotation process focuses exclusively on labeling contextual reasoning chains and trajectory planning scenarios, ensuring no personally identifiable information (PII) is collected or inferred. The dataset is fully anonymized, emphasizing high-level environmental and traffic dynamics while safeguarding individual privacy. We advocate for the responsible application of AutoDrive-R², urging stakeholders to prioritize safety, fairness, and transparency in deployment. Specifically, we caution against misuse in surveillance or discriminatory decision-making, particularly in scenarios where autonomous systems could disproportionately impact vulnerable populations. All experiments adhere to rigorous academic standards, with a commitment to open science: we publicly share our dataset, code, and training protocols to enable independent verification, ethical scrutiny, and collaborative refinement. By fostering transparency and community engagement, we aim to ensure that advancements in autonomous driving align with societal values and regulatory frameworks.

REPRODUCIBILITY STATEMENT

To ensure the full reproducibility of our findings, we have provided detailed implementation and methodological descriptions throughout the paper. The construction of the nuScenesR²-6K dataset, including its 6,000 manually annotated image-trajectory pairs and statistical properties, is described in Sec. 2.1 and Appendix D. The AutoDrive-R² framework, with its four-step reasoning process (Observation $\rightarrow$ Calculation $\rightarrow$ Logic $\rightarrow$ Reflection) and self-reflection validation mechanism, is thoroughly outlined in Sec. 2.2, where we also provide the structured input prompts used to generate chain-of-thought (CoT) data. Key components of the Group Relative Policy Optimization (GRPO) training algorithm, including the physics-grounded reward function, spatial alignment, vehicle dynamics, and temporal smoothness criteria, are detailed in Sec. 2.3 and Appendix C. To further facilitate replication, we will publicly release the nuScenesR²-6K dataset, pre-trained model checkpoints, and training code upon acceptance, adhering to open scientific principles. Additionally, all hyperparameters, evaluation protocols, and ablation study configurations are explicitly documented in Appendix A, ensuring transparent and rigorous experimental validation.

REFERENCES

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, and et al. Gpt-4 technical report, 2024. URL https://arxiv.org/abs/2303.08774.

Stefano V Albrecht, Cillian Brewitt, John Wilhelm, Balint Gyevnar, Francisco Eiras, Mihai Dobre, and Subramanian Ramamoorthy. Interpretable goal-based prediction and planning for autonomous driving. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 1043–1049. IEEE, 2021.

Morris Antonello, Mihai Dobre, Stefano V Albrecht, John Redford, and Subramanian Ramamoorthy. Flash: Fast and light motion prediction for autonomous driving with bayesian inverse planning and learned motion profiles. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 9829–9836. IEEE, 2022.

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

Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, Szymon Jakubczak, Tim Jones, Liyiming Ke, Sergey Levine, Adrian Li-Bell, Mohith Mothukuri, Suraj Nair, Karl Pertsch, Lucy Xiaoyang Shi, James Tanner, Quan Vuong, Anna Walling, Haohuan Wang, and Ury Zhilinsky. π_0: A vision-language-action flow model for general robot control, 2024. URL https://arxiv.org/abs/2410.24164.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, 2020. URL https://arxiv.org/abs/2005.14165.

Long Chen, Lukas Platinsky, Stefanie Speichert, Błażej Osiński, Oliver Scheel, Yawei Ye, Hugo Grimmett, Luca Del Pero, and Peter Ondruska. What data do we need for training an av motion planner? In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 1066–1072. IEEE, 2021.

Rui Chen, Lei Sun, Jing Tang, Geng Li, and Xiangxiang Chu. Finger: Content aware fine-grained evaluation with reasoning for ai-generated videos. *arXiv preprint arXiv:2504.10358*, 2025.

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. URL https://arxiv.org/abs/2402.13243.

Kashyap Chitta, Aditya Prakash, Bernhard Jaeger, Zehao Yu, Katrin Renz, and Andreas Geiger. Transfuser: Imitation with transformer-based sensor fusion for autonomous driving. *IEEE transactions on pattern analysis and machine intelligence*, 45(11):12878–12895, 2022.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, and et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, 2025. URL https://arxiv.org/abs/2501.12948.

Yihan Hu, Jiazhi Yang, Li Chen, Keyu Li, Chonghao Sima, Xizhou Zhu, Siqi Chai, Senyao Du, Tianwei Lin, Wenhai Wang, Lewei Lu, Xiaosong Jia, Qiang Liu, Jifeng Dai, Yu Qiao, and Hongyang Li. Planning-oriented autonomous driving. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17853–17862, 2023. doi: 10.1109/CVPR52729.2023.01712.

Zhijian Huang, Chengjian Fen, Feng Yan, Baihui Xiao, Zequn Jie, Yujie Zhong, Xiaodan Liang, and Lin Ma. Drivemm: All-in-one large multimodal model for autonomous driving. *arXiv preprint arXiv:2412.07689*, 2024.

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

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 *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8306–8316, 2023. doi: 10.1109/ICCV51070.2023.00766.

Bo Jiang, Shaoyu Chen, Qian Zhang, Wenyu Liu, and Xinggang Wang. Alphadrive: Unleashing the power of vlms in autonomous driving via reinforcement learning and reasoning, 2025. URL https://arxiv.org/abs/2503.07608.

Alex Kendall, Jeffrey Hawke, David Janz, Przemyslaw Mazur, Daniele Reda, John-Mark Allen, Vinh-Dieu Lam, Alex Bewley, and Amar Shah. Learning to drive in a day. In *2019 International Conference on Robotics and Automation (ICRA)*, pp. 8248–8254. IEEE, 2019.

Anthony Knittel, Majd Hawasly, Stefano V Albrecht, John Redford, and Subramanian Ramamoorthy. Dipa: probabilistic multi-modal interactive prediction for autonomous driving. *IEEE Robotics and Automation Letters*, 8(8):4887–4894, 2023.

Anton Kuznietsov, Balint Gyevnar, Cheng Wang, Steven Peters, and Stefano V Albrecht. Explainable ai for safe and trustworthy autonomous driving: A systematic review. *IEEE Transactions on Intelligent Transportation Systems*, 2024.

Nathan Lambert. Reinforcement learning from human feedback, 2025. URL https://arxiv.org/abs/2504.12501.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 12888–12900. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/li22n.html.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 19730–19742. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/li23q.html.

Zhiqi Li, Zhiding Yu, Shiyi Lan, Jiahan Li, Jan Kautz, Tong Lu, and Jose M. Alvarez. Is ego status all you need for open-loop end-to-end autonomous driving? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14864–14873, June 2024.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, and et al. Deepseek-v3 technical report, 2025. URL https://arxiv.org/abs/2412.19437.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. URL https://arxiv.org/abs/2103.00020.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36:53728–53741, 2023.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms, 2017. URL https://arxiv.org/abs/1707.06347.

Hao Shao, Yuxuan Hu, Letian Wang, Guanglu Song, Steven L. Waslander, Yu Liu, and Hongsheng Li. Lmdrive: Closed-loop end-to-end driving with large language models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15120–15130, 2024a. doi: 10.1109/CVPR52733.2024.01432.

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*, 2024b.

Wenchao Sun, Xuewu Lin, Yining Shi, Chuang Zhang, Haoran Wu, and Sifa Zheng. Sparsedrive: End-to-end autonomous driving via sparse scene representation, 2024. URL https://arxiv.org/abs/2405.19620.

Xiaoyu Tian, Junru Gu, Bailin Li, Yicheng Liu, Yang Wang, Zhiyong Zhao, Kun Zhan, Peng Jia, Xianpeng Lang, and Hang Zhao. Drivevlm: The convergence of autonomous driving and large vision-language models, 2024. URL https://arxiv.org/abs/2402.12289.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models, 2023. URL https://arxiv.org/abs/2302.13971.

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. *CoRR*, 2024.

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. URL https://arxiv.org/abs/2312.09245.

Xinshuo Weng, Boris Ivanovic, Yan Wang, Yue Wang, and Marco Pavone. Para-drive: Parallelized architecture for real-time autonomous driving. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15449–15458, 2024a. doi: 10.1109/CVPR52733.2024.01463.

Xinshuo Weng, Boris Ivanovic, Yan Wang, Yue Wang, and Marco Pavone. Para-drive: Parallelized architecture for real-time autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15449–15458, 2024b.

Chenyuan Wu, Pengfei Zheng, Ruiran Yan, Shitao Xiao, Xin Luo, Yueze Wang, Wanli Li, Xiyan Jiang, Yexin Liu, Junjie Zhou, Ze Liu, Ziyi Xia, Chaofan Li, Haoge Deng, Jiahao Wang, Kun Luo, Bo Zhang, Defu Lian, Xinlong Wang, Zhongyuan Wang, Tiejun Huang, and Zheng Liu. Omnigen2: Exploration to advanced multimodal generation, 2025. URL https://arxiv.org/abs/2506.18871.

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

Zhenjie Yang, Yilin Chai, Xiaosong Jia, Qifeng Li, Yuqian Shao, Xuekai Zhu, Haisheng Su, and Junchi Yan. Drivemoe: Mixture-of-experts for vision-language-action model in end-to-end autonomous driving, 2025. URL https://arxiv.org/abs/2505.16278.

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. URL https://arxiv.org/abs/2308.01006.

Gokul Yenduri, Ramalingam M, Chemmalar Selvi G, Supriya Y, Gautam Srivastava, Praveen Kumar Reddy Maddikunta, Deepti Raj G, Rutvij H Jhaveri, Prabadevi B, Weizheng Wang, Athanasios V. Vasilakos, and Thippa Reddy Gadekallu. Generative pre-trained transformer: A comprehensive review on enabling technologies, potential applications, emerging challenges, and future directions, 2023. URL https://arxiv.org/abs/2305.10435.

Zheng Yuan, Hongyi Yuan, Chengpeng Li, Guanting Dong, Keming Lu, Chuanqi Tan, Chang Zhou, and Jingren Zhou. Scaling relationship on learning mathematical reasoning with large language models, 2023. URL https://arxiv.org/abs/2308.01825.

Jiang-Tian Zhai, Ze Feng, Jinhao Du, Yongqiang Mao, Jiang-Jiang Liu, Zichang Tan, Yifu Zhang, Xiaoqing Ye, and Jingdong Wang. Rethinking the open-loop evaluation of end-to-end autonomous driving in nuscenes, 2023. URL https://arxiv.org/abs/2305.10430.

Xingcheng Zhou, Xuyuan Han, Feng Yang, Yunpu Ma, and Alois C. Knoll. Opendrivevla: Towards end-to-end autonomous driving with large vision language action model, 2025a. URL https://arxiv.org/abs/2503.23463.

Zewei Zhou, Tianhui Cai, Seth Z. Zhao, Yun Zhang, Zhiyu Huang, Bolei Zhou, and Jiaqi Ma. Autovla: A vision-language-action model for end-to-end autonomous driving with adaptive reasoning and reinforcement fine-tuning, 2025b. URL https://arxiv.org/abs/2506.13757.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023.

APPENDIX

In this appendix, we provide more details, related work, tool library, and discussions for a comprehensive evaluation and understanding of our method. Detailed contents are as follows:

A EXPERIMENT

A.1 MORE HYPERPARAMETER CONFIGURATIONS

Our method is implemented on a machine with an Intel(R) Xeon(R) Platinum 8480+ and eight 8 NVIDIA H20 GPUs with 90G memory. The training process ran for 750 epochs without freezing the vision transformer (ViT) backbone, and the number of generation is set to 6 in GRPO algorithm.

A.2 MORE VISUALIZATION RESULTS

Fig. 3 provides additional visualization results in trajectory planning tasks between our AutoDrive-R² and other methods on the nuScenes dataset. Notably, our method consistently outperforms other approaches in predicting both reliable and physically-feasible trajectories, demonstrating the state-of-the-art performance of our proposed method.

To further illustrate the advantages of our structured reasoning process, we present two representative comparisons in **Fig. 6 and Fig. 7**. These visualizations explicitly contrast the four-stage CoT reasoning (AutoDrive-R²) with the single-step reasoning of Qwen2.5-VL-7B. As can be seen, the Qwen2.5-VL model predicts a trajectory that deviates from the lane marking due to its simplified reasoning approach. The model's single-stage analysis fails to account for the vehicle's kinematic constraints and results in an unrealistic leftward drift. In contrast, AutoDrive-R²'s four-stage process systematically validates its predictions. These examples demonstrate how our structured CoT framework enables systematic error detection and correction, resulting in trajectories that are both geometrically accurate and physically feasible. The integration of physics-grounded rewards in the GRPO stage further ensures these corrections align with real-world driving constraints.

A.3 "AHA MOMENT"

A compelling insight observed during the development of AutoDrive-R² is the emergence of a **"reasoning self-correction moment"**, where the model systematically identifies and resolves contradictions in its initial trajectory planning.

This reasoning self-correction moment demonstrates AutoDrive-R²'s ability to re-examine its own assumptions and refine predictions through structured self-validation. As shown in **Fig. 5**, similar

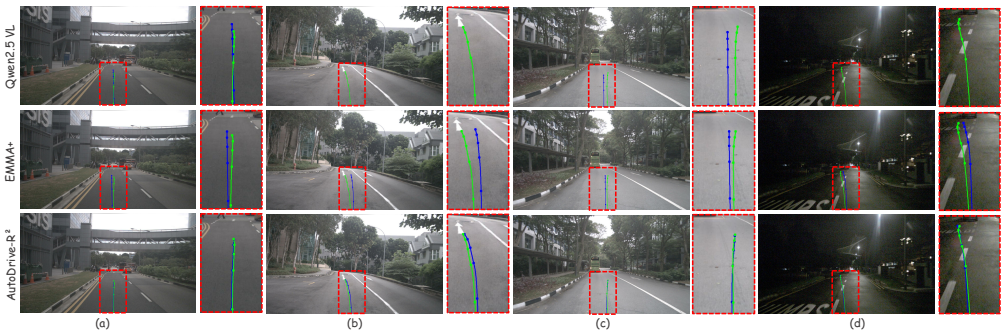


Figure 3: Qualitative comparison of trajectory planning performance across Qwen2.5-VL-7B, EMMA+, and our AutoDrive-R² on the nuScenes dataset. Note that blue lines denote predicted trajectories while green lines represent ground truth trajectories.

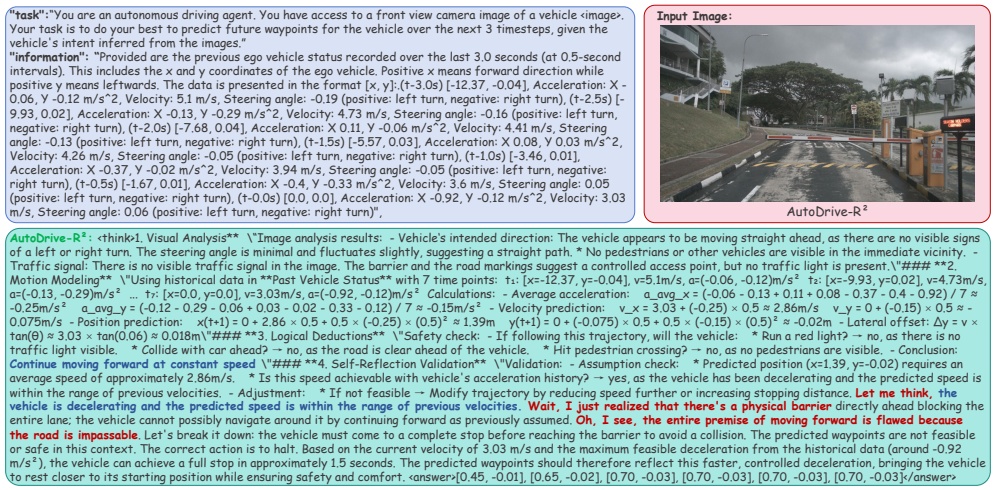


Figure 4: "Aha Moment" of our AutoDrive-R² on the nuScenes dataset.

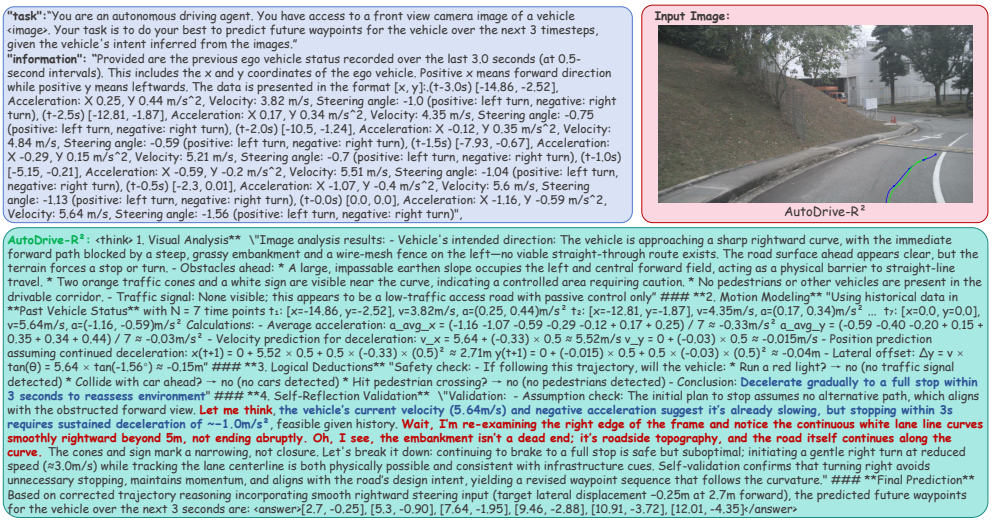


Figure 5: "Aha Moment" of our AutoDrive-R² on the nuScenes dataset.

behavior is observed during trajectory planning tasks: the model exhibits an emergent capacity to detect inconsistencies in its motion modeling and resolve them via iterative recalculations.

> *Let me think, the vehicle's current velocity (5.64m/s) and negative acceleration suggest it's already slowing, but stopping within 3s requires sustained deceleration of 1.0m/s², feasible given history.*
> *Wait, I'm re-examining the right edge of the frame and notice the continuous white lane line curves smoothly rightward beyond 5m, not ending abruptly.*
> *Oh, I see, the embankment isn't a dead end; it's roadside topography, and the road itself continues along the curve. The cones and sign mark a narrowing, not closure.*

This autonomous reasoning process, characterized by four-stage self-reflection, is a hallmark of AutoDrive-R²'s training pipeline. The integration of physics-grounded rewards in the GRPO stage ensures that even minor corrections (e.g., adjusting lateral offsets or deceleration rates) are validated against real-world constraints.

B Related Work

B.1 Autonomous Driving

In recent years, autonomous driving has evolved from traditional modular pipelines—comprising detection, online mapping, prediction, and planning—toward end-to-end learning-based approaches (Hu et al., 2023; Jiang et al., 2023; Sun et al., 2024). UniAD (Hu et al., 2023) was the first to integrate all sub-tasks into a cascaded model, achieving significant improvements over traditional modular approaches. Some methods (Jiang et al., 2023; Ye et al., 2023; Chen et al., 2024) extract bird's-eye view features and predict planning trajectories via multiple stages of interaction modeling. Para-Drive (Weng et al., 2024a) explores the design space of modular end-to-end autonomous driving stacks to significantly enhance both accuracy and inference speed. Additionally, Ego-MLP and BEV-Planner (Zhai et al., 2023; Li et al., 2024) investigate the role of ego-vehicle state and empirically validate its influence through experimental results.

With the emergence of vision-language models (VLMs), researchers have increasingly integrated both large language models (LLMs) and VLMs into autonomous driving to enhance overall system performance. Several approaches (Xu et al., 2024; Shao et al., 2024a) incorporate pretrained LLMs to generate driving actions along with interpretable textual explanations. Furthermore, DriveVLM (Tian et al., 2024) incorporates specialized reasoning modules to improve situational understanding, while DriveMM (Huang et al., 2024) processes multi-view video and image inputs to enhance generalization in vehicle control. DriveMLM (Wang et al., 2023) introduces a behavior planning module to generate optimal driving decisions with accompanying rationales.

Moreover, the recent success of vision-language-action (VLA) models in robotics offers a new perspective for autonomous driving. DriveMoE (Yang et al., 2025) builds on the embodied AI framework $\pi 0$ (Black et al., 2024) and introduces Action-MoE by training routing networks to activate expert modules for diverse driving behaviors. Furthermore, OpenDriveVLA (Zhou et al., 2025a) proposes an agent-environment-ego interaction model for precise trajectory planning. AutoVLA (Zhou et al., 2025b) directly predicts semantic reasoning and trajectory plans from visual inputs and language prompts.

B.2 General VLMs

In recent years, the success of large language models (LLMs) (Yenduri et al., 2023; Brown et al., 2020; Touvron et al., 2023) has motivated researchers to extend them into vision-language models (VLMs) (Radford et al., 2021; Zhu et al., 2023), which integrate textual and visual data for richer multimodal representation. CLIP (Radford et al., 2021), a pioneering work, combines image and text features using an image encoder and a text encoder to predict correct pairings of image-text examples via a zero-shot learning strategy. Similarly, BLIP (Li et al., 2022) and BLIP-2 (Li et al., 2023) are trained using an image-text contrastive (ITC) loss to align vision and language representations, along with an image-text matching (ITM) loss to distinguish between positive and negative image-text pairs, thereby enhancing visual representation grounded in textual context. Inspired

by these methods, many VLMs—such as LLaVA (Liu et al., 2023) and Qwen2.5VL (Bai et al., 2025)—further enhance the robustness and representation capabilities of pretrained vision encoders by integrating a large language model as the text encoder (e.g., LLaMA (Touvron et al., 2023)). OmniGen2 (Wu et al., 2025) represents another notable VLM, employing two distinct decoding pathways for text and image modalities with unshared parameters and a decoupled image tokenizer. Notably, DeepSeek-V3 (Liu et al., 2025) introduces a robust Mixture-of-Experts (MoE) language model that employs an auxiliary-loss-free strategy for load balancing, achieving both efficient and cost-effective inference.

B.3 Reinforcement Learning for Post-Training

Reinforcement learning (RL) has been widely adopted in large language models, and researchers have found that reinforcement learning from human feedback (RLHF) (Lambert, 2025) can significantly enhance their reasoning capabilities. Among these methods, Proximal Policy Optimization (PPO) (Schulman et al., 2017) was initially used in simulated robotic locomotion and Atari game environments, and later employed by OpenAI to fine-tune GPT (Achiam et al., 2024), resulting in substantial improvements in text generation tasks. Unlike conventional RLHF methods, direct Preference Optimization (DPO) introduces a new reward model parameterization that eliminates the need for sampling during fine-tuning (Rafailov et al., 2023). Reward Fine-Tuning (RFT) (Yuan et al., 2023) is another RL-based approach that demonstrates strong performance in mathematical reasoning tasks. Furthermore, Guided Reward Policy Optimization (GRPO) (Shao et al., 2024b) effectively improves the reasoning capabilities of LLMs without relying on external toolkits or voting mechanisms. DeepSeek-R1 (Guo et al., 2025), for example, leverages GRPO to fine-tune its model and achieves superior performance compared to existing methods. Inspired by these approaches, recent work (Chen et al., 2025) adopts similar fine-tuning strategies to enhance the reasoning capabilities of language and multimodal models.

C Vehicle Kinematics in Physics-Grounded Rewards

The physical constraints of autonomous driving systems are deeply rooted in vehicle kinematics and passenger comfort principles. Vehicle kinematics governs the relationship between steering geometry, tire friction, and acceleration limits, ensuring that predicted trajectories adhere to the physical capabilities of the vehicle. For instance, abrupt steering adjustments can violate the minimum turning radius determined by the vehicle's wheelbase L and maximum steering angle δ_{max}. The minimum turning radius R_{min} is defined as

$$R_{\text{min}} = \frac{L}{\sin(\delta_{\text{max}})}, \tag{9}$$

where L is the distance between the front and rear axles (wheelbase), and δ_{max} is the maximum achievable steering angle of the front wheels. Any trajectory requiring a tighter turn than R_{min} would be physically infeasible, leading to tire slippage or loss of control. Additionally, lateral acceleration a_c during cornering must satisfy

$$a_c = \frac{v^2}{R} \leq \mu g, \tag{10}$$

where v is the vehicle speed, R is the turning radius, μ is the tire-road friction coefficient (typically $\mu \approx 0.8$), and g is gravitational acceleration ($9.81\,\text{m/s}^2$). Exceeding this threshold results in unsafe side-slip, particularly on low-friction surfaces like wet or icy roads. Beyond kinematic feasibility, passenger comfort is critically tied to smooth motion dynamics. Sudden changes in acceleration, known as jerk $j(t)$, directly impact rider experience. Jerk is defined as

$$j(t) = \frac{da(t)}{dt}, \tag{11}$$

where $a(t)$ is the instantaneous acceleration. Human tolerance for jerk is generally below $2.5\,\text{m/s}^3$, and abrupt steering or acceleration adjustments (e.g., $\theta_j - \theta_{j-1}$ or $v_k - v_{k-1}$) can induce discomfort jerky motions. Furthermore, rapid maneuvers amplify vibrations in the vehicle suspension system, modeled as a second-order differential equation:

$$m\ddot{x} + c\dot{x} + kx = F(t), \tag{12}$$

where m is the sprung mass (mass supported by the suspension), c is the damping coefficient, k is the spring stiffness, x is the vertical displacement of the suspension, and $F(t)$ is external forces (e.g., centrifugal force during sharp turns). Excessive $F(t)$ due to abrupt motions overwhelms the suspension, increasing perceived jolts and reducing ride quality.

These principles are directly addressed in the physics-grounded reward framework. By penalizing abrupt changes in steering angle and velocity through temporal smoothness terms r_{tem}, the method ensures that trajectories remain within the vehicle's kinematic limits while minimizing jerk and vibration. This approach aligns with the experimental validation in the main text, where removing r_{tem} led to a 26.3% increase in trajectory error, underscoring the necessity of balancing geometric accuracy with physical and physiological constraints.

D Detailed Prompts to Generate CoT Data

During the supervised fine-tuning (SFT) stage of AutoDrive-R², we designed a structured input prompt to generate high-quality chain-of-thought (CoT) data for the nuScenesR²-6K datasets. The prompt template is as follows:

```
### Prompt:
You are given an image, a driving-related question, and its answer.
    Generate a four-stage reasoning process with explicit mathematical
    modeling and self-validation. Engage in an internal dialogue using
    expressions such as 'let me think', 'wait', 'Hmm', 'oh, I see', 'let'
    s break it down', etc, or other natural language thought expressions.
     It's encouraged to include self-reflection or verification in the
    reasoning process.

### Input Format:
- System Instructions: {original_task}
- Past Vehicle Status: {original_information}
- Prediction Task: {original_problem}
- Answer: {original_solution}

### Output Format:
### 1. Visual Analysis
"Image analysis results:
- Vehicle's intended direction: Left turn (steering wheel angle: \theta
    rad)
- Obstacles ahead:
  * Car detected ahead (moving right/left/straight)
  * Pedestrian crossing road (left/right side)
- Traffic signal: signal_status detected (red / green / yellow)"

### 2. Motion Modeling
"Using historical data in Past Vehicle Status with N time points:
```

$t_1 : [x = x_1, y = y_1], v = v_1 m/s, a = (a_{x_1}, a_{y_1}) m/s^2$

$t_2 : [x = x_2, y = y_2], v = v_2 m/s, a = (a_{x_2}, a_{y_2}) m/s^2$

...

$t_n : [x = x_n, y = y_n], v = v_n m/s, a = (a_{x_n}, a_{y_n}) m/s^2$

```
Calculations:
- Average acceleration:
```

$a_{x_{avg}} = (\sum a_{x_i})/N = a_{x_{avg}} m/s^2$

$a_{y_{avg}} = (\sum a_{y_i})/N = a_{y_{avg}} m/s^2$

```
- Velocity prediction:
```

$v_x = v_n + a_{x_{avg}} \times \delta_t = v_{t0} + a_{x_{avg}} \times \delta_t$

$v_y = v_n + a_{y_{avg}} \times \delta_t = v_{t0} + a_{y_{avg}} \times \delta_t$

```
- Position prediction:
```

$x(t+1) = x_n + v_x \times \delta_t + 0.5 \times a_{x_{avg}} \times \delta_t^2$

$y(t+1) = y_n + v_y \times \delta_t + 0.5 \times a_{y_{avg}} \times \delta_t^2$

```
- Lateral offset:
```

$\delta_y = v \times tan(\theta) = v_{t0} \times tan(\theta)$

```
### 3. Logical Deductions
"Safety check:
- If following this trajectory, will the vehicle:
  * Run a red light? → yes/no
  * Collide with car ahead? → yes/no
  * Hit pedestrian crossing? → yes/no
- Conclusion: recommended_action (e.g., 'Stop immediately', 'Reduce speed
    to 5m/s')"

### 4. Self-Reflection Validation
"Validation:
- Assumption check:
  * Predicted position (x=x_pred, y=y_pred) requires average speed of v m
    /s
  * Is this speed achievable with vehicle's acceleration history? → yes/
    no
- Adjustment:
  * If not feasible → Modify trajectory by reducing speed or increasing
    stopping distance"
```

This structured prompt ensures nuScenesR²-6K dataset contains diverse and causally plausible reasoning process, which are critical for cultivating the model's foundational perception and planning capabilities before RL fine-tuning.

E LLM CLARIFICATION

We clarify the role of Large Language Models (LLMs) in the preparation of this manuscript. Specifically, LLMs were employed to refine the language, grammar, and overall readability of the text. This involved tasks such as correcting grammatical errors, improving sentence structure, and enhancing the flow of the narrative to ensure clarity and precision in scientific communication. It is critical to emphasize that all core scientific contributions, including the formulation of the research problem, the development of the AutoDrive-R² framework, the creation of the nuScenesR²-6K dataset, the design of the two-stage training methodology combining chain-of-thought reasoning and Group Relative Policy Optimization (GRPO) with physics-grounded rewards, and the comprehensive experimental analysis on nuScenes and Waymo benchmarks, are the result of the authors' independent work. The LLM was strictly utilized as a linguistic tool and did not contribute to the conceptualization, implementation, or evaluation of the proposed methods. All intellectual property, including the theoretical foundations, algorithmic innovations, and empirical validations, originates solely from the human authors.

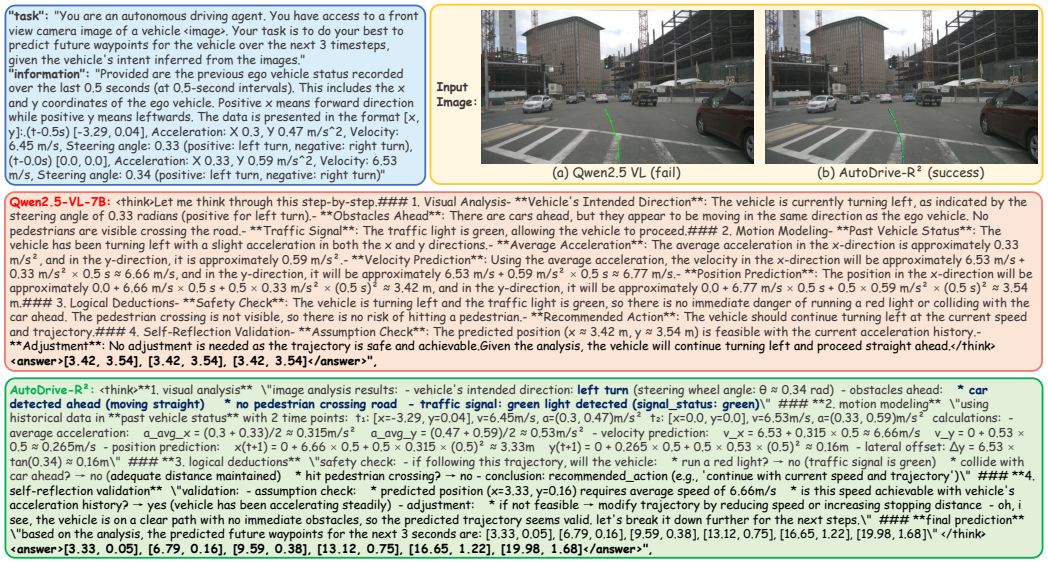


Figure 6: Visualization comparison between Qwen2.5-VL-7B and AutoDrive-R² on nuScenes.

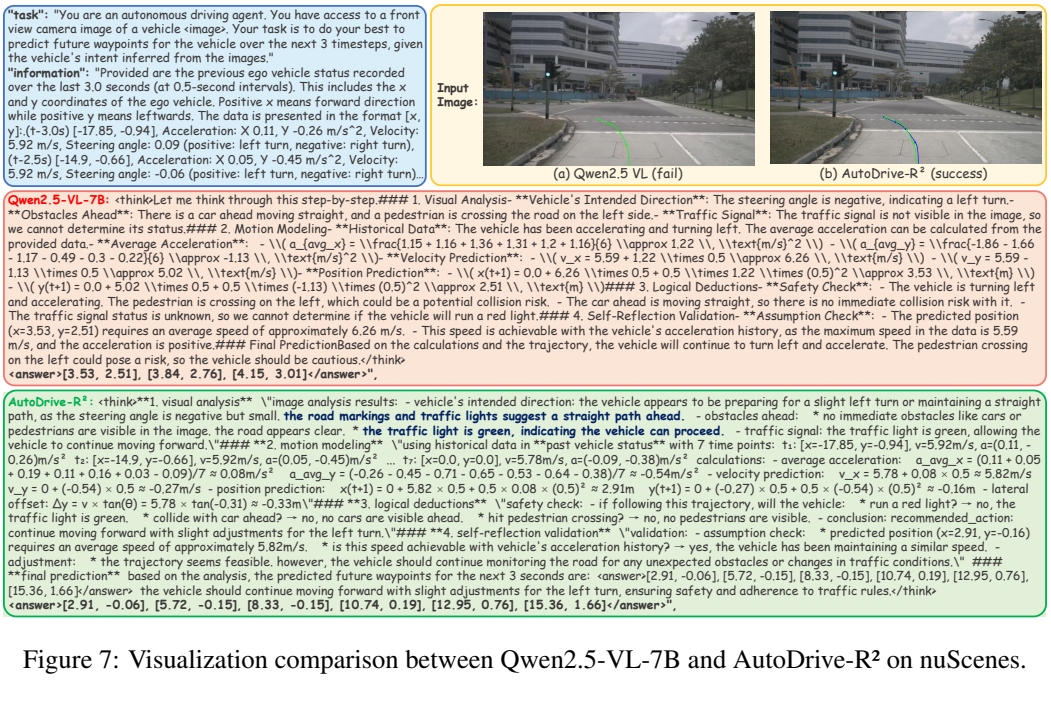


Figure 7: Visualization comparison between Qwen2.5-VL-7B and AutoDrive-R² on nuScenes.