CritiqueDriveVLM: From Verifier-Guided Reinforcement Learning to Latent Thought Distillation for Autonomous Driving

Zhaohong Liu[1,3], Hao Ye[1,3], Xianlin Zhang[2,3], and Mengshi Qi[1,3(✉)]

[1] State Key Laboratory of Networking and Switching Technology
[2] School of Digital Media & Design Arts
[3] Beijing University of Posts and Telecommunications, Beijing, China
{liuzhaoh, haoye, zxlin, qms}@bupt.edu.cn

Abstract. End-to-end Vision-Language Models (VLMs) show immense potential in autonomous driving. However, standard Supervised Fine-Tuning (SFT) often suffers from reasoning hallucinations and conservative biases. While traditional tool-augmented frameworks and Chain-of-Thought (CoT) approaches mitigate these issues, they incur exorbitant token consumption and unacceptable latency, rendering real-time deployment impractical. To resolve this reliability-efficiency trade-off, we propose **CritiqueDriveVLM**, a novel unified three-stage framework internalizing reasoning directly into the VLM. First, we introduce Critique-Driven Multi-Turn Reinforcement Learning (RL) guided by a multi-dimensional verifier. By providing granular scalar feedback and a multi-turn penalty, we force the policy to internalize logical deduction, cultivating a robust System-2 Teacher that achieves high accuracy without fragile external tools. Subsequently, we propose Latent Thought Distillation to overcome the latency bottleneck. By aligning the Student's latent representations with the Teacher's fully converged reasoning states, we compress deep logical capabilities into a fast, CoT-free System-1 Student. Extensive experiments on the widely-used DriveLMM-o1 benchmark demonstrate remarkable improvements. Compared to the base model, our tool-free Teacher significantly boosts Multiple Choice Quality (MCQ) from 55.54% to a state-of-the-art 76.54%. Crucially, our distilled Student preserves competitive reasoning depth while drastically minimizing generation length to an average of merely 28 tokens. This slashes inference latency by 88% (from 3482 ms to 416 ms), paving a highly robust pathway for low-latency autonomous driving. Our source code is available at https://github.com/MICLAB-BUPT/CritiqueDriveVLM.

Keywords: Vision-Language Models · Autonomous Driving · Reinforcement Learning · Knowledge Distillation

1 Introduction

The integration of Vision-Language Models (VLMs) [1, 2, 4, 28, 43] has catalyzed a profound paradigm shift in autonomous driving, transitioning the field from

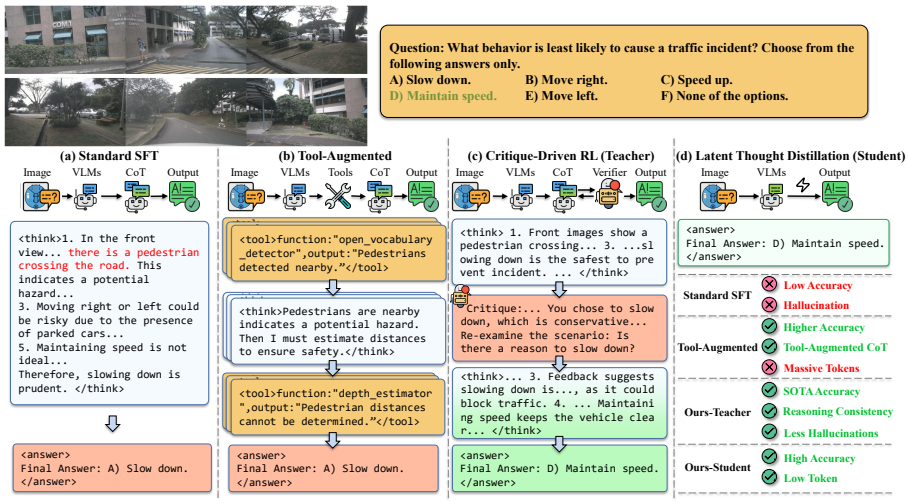

Fig. 1: Paradigm comparison of VLM-based autonomous driving. (a) Standard SFT is prone to reasoning hallucinations and conservative biases. (b) Tool-Augmented methods suffer from brittle external APIs and high latency. (c) Critique-Driven RL (Teacher) internalizes deep logic without relying on external tools. (d) Latent Thought Distillation (Student) enables instant, CoT-free execution, eliminating the overhead of explicit reasoning tokens.

traditional, modular perception-planning pipelines [8, 9, 23, 25] toward end-to-end holistic reasoning frameworks [16, 19]. By leveraging extensive world knowledge and powerful visual-semantic alignment, VLMs demonstrate remarkable potential in interpreting complex traffic scenes and understanding nuanced dynamic interactions. However, deploying these models directly into safety-critical environments remains an open challenge. Current methodologies predominantly rely on standard Supervised Fine-Tuning (SFT) [47] to align pre-trained VLMs with driving-specific instructions. Unfortunately, standard SFT merely mimics the surface-level statistical distribution of human driving trajectories without genuinely internalizing deep logical deduction capabilities. Consequently, in complex dynamic scenarios, SFT models frequently suffer from severe reasoning hallucinations and conservative biases. As illustrated in Figure 1(a), a standard SFT model might easily misinterpret ambiguous visual cues, leading to suboptimal or overly cautious decisions, such as unnecessary deceleration, which disrupts traffic flow and fundamentally compromises safety.

To mitigate these inherent reasoning deficiencies, recent state-of-the-art methodologies have increasingly turned to augmenting VLMs with Chain-of-Thought (CoT) [57] processes and explicit external tool calls. Frameworks such as AgentThink [42] and OmniDrive-R1 [62] attempt to actively ground visual elements and verify intermediate logic by interleaving multi-modal reasoning with dynamic Application Programming Interface (API) calls, such as open-

vocabulary detectors or depth estimators. However, as depicted in Figure 1(b), this tool-augmented paradigm fundamentally compromises the robustness of the end-to-end architecture. It introduces a fragile dependency pipeline where a single point of failure or a noisy output from an external module can cascade, causing the entire decision-making process to collapse or produce hazardous directives. Furthermore, generating hundreds of explicit reasoning tokens imposes prohibitive computational overhead, rendering these prolonged System-2 approaches impractical for real-time vehicular deployment.

To overcome the fragility of tool-augmented pipelines, we introduce a new Critique-Driven Multi-Turn Reinforcement Learning (RL) paradigm in the second stage of our CritiqueDriveVLM framework. Instead of relying on brittle external perception APIs, we construct a frozen Multi-Dimensional Verifier that strictly evaluates the model's reasoning trajectories across perception, logic, and safety dimensions. As illustrated in Figure 1(c), this verifier acts as a logical guide during inference, providing targeted natural language critiques that allow the Teacher model to engage in a multi-turn interactive loop, iteratively refining its initial thoughts and correcting hallucinations. Concurrently, during the Group Relative Policy Optimization (GRPO) [49] training phase, the verifier provides fine-grained scalar rewards, while a step-decay multi-turn penalty explicitly forces the policy to internalize the logical deduction process. This cultivates a highly reliable Teacher model that achieves profound System-2 reasoning and robust decision-making through verifier-guided refinement.

While our Stage-2 Teacher achieves state-of-the-art accuracy, the heavy computational burden of explicit CoT reasoning limits its real-time applicability. Conventional knowledge distillation [13] typically forces a lightweight student either to mimic the teacher's verbose CoT tokens, failing to resolve the latency bottleneck, or to directly replicate the final answer, severely truncating the required logical depth. To bridge this gap, we introduce a novel Latent Thought Distillation strategy in the third stage. Unlike tool-augmented methods that remain heavily bottlenecked by explicit API calls latency and fragile parsing pipelines, our proposed distillation operates entirely within the model's internal latent space. Rather than imitating surface-level text, we directly align the Student's latent representations with the Teacher's fully converged hidden states at the end of its reasoning trajectory, specifically at the final `</think>` token. As depicted in Figure 1(d), this novel objective effectively compresses the complex reasoning capabilities of the System-2 Teacher into a fast, CoT-free System-1 Student. The distilled Student completely bypasses both the external verifier and the intermediate text generation, predicting the safe driving action with ultra-low latency, thereby achieving an optimal balance between deep reasoning rigor and low-latency inference efficiency.

The main contributions of this work can be summarized as follows:

1. We propose **CritiqueDriveVLM**, a novel three-stage framework specifically designed to bridge the gap between deep System-2 logic and low-latency System-1 execution for autonomous driving VLMs.

2. We introduce **Critique-Driven Multi-Turn RL** equipped with a multi-dimensional verifier and step-decay penalty, successfully suppressing visual hallucinations without relying on fragile external tools.
3. We develop **Latent Thought Distillation** to align the student's hidden representations with the teacher's fully converged reasoning states, enabling highly accurate CoT-free inference.
4. Extensive experiments on the **DriveLMM-o1** [17] benchmark demonstrate the superiority of our approach. Our tool-free Teacher model achieves state-of-the-art reasoning accuracy, while our distilled Student model drastically slashes inference latency to low-latency levels, paving a practical pathway for robust autonomous driving.

2 Related Work

2.1 VLMs in Autonomous Driving

The integration of VLMs has shifted autonomous driving from modular [7, 8] pipelines to end-to-end holistic reasoning [16, 19, 26, 30, 55, 64]. Pioneering works such as DriveVLM [32, 54] demonstrate the potential of VLMs in complex scene understanding and hierarchical planning. To systematically evaluate and advance these capabilities, benchmarks such as DriveLMM-o1 [17, 52] have been introduced, providing specialized datasets that emphasize step-wise visual reasoning and deep logical inference in complex dynamic scenarios. To mitigate reasoning hallucinations, recent state-of-the-art approaches rely on tool-augmented architectures and prolonged CoT. For instance, AgentThink [42] integrates CoT with dynamic tool calls, while OmniDrive-R1 [62] employs an interleaved Multi-modal CoT (iMCoT) mechanism for fine-grained active visual grounding. However, these methods incur exorbitant token consumption and latency bottlenecks [60]. In contrast, our proposed CritiqueDriveVLM eliminates dependencies on external APIs and explicit CoT generation, achieving robust low-latency inference.

2.2 Reinforcement Learning with Verifiable Rewards

Reinforcement Learning with Verifiable Rewards (RLVR) has emerged as a powerful paradigm to unlock the reasoning potential of Large Language Models [27, 33]. Pioneering works such as DeepSeekMath [49] and DeepSeekMathV2 [48] demonstrate that RL, particularly through algorithms such as GRPO, can significantly boost logical performance in structured domains by optimizing against objective, rule-based rewards. DeepSeek-R1 [11] further showcases how large-scale RLVR elicits emergent self-correction and sophisticated reasoning trajectories. Recently, AlphaDrive [20] introduces a two-stage GRPO-based RL framework tailored for autonomous driving, demonstrating that RL can effectively elicit emergent multimodal planning capabilities beyond standard SFT [5, 21, 45, 63]. However, applying RLVR to complex, open-ended driving scenarios often suffers from unreliable or sparse reward signals, inevitably leading to severe hallucinations [24] during the CoT process. To address this, our framework introduces

a multi-dimensional verifier to provide granular feedback during RL, effectively suppressing CoT hallucinations and establishing a highly reliable Teacher model capable of critique-driven refinement for subsequent distillation.

2.3 Knowledge Distillation

Knowledge Distillation [13, 34–41, 44] is fundamentally designed to transfer capabilities from a cumbersome teacher model to a lightweight student. Early alignment methods, such as Self-Instruct [56], primarily relied on bootstrapping instruction-following data to mimic output distributions. Subsequent advancements, including Distilling Step-by-Step [14] and MiniLLM [10], explicitly distilled CoT rationales to enhance generation quality in smaller models. Recently, researchers have explored internalizing slow System 2 reasoning into fast System 1 intuition [6, 12, 22, 50, 61], while the distillation phase of DeepSeek-R1 [11] demonstrates that smaller models can successfully inherit complex reasoning capabilities from large RL-trained models. However, existing distillation paradigms typically force the student model to either mimic the teacher's explicit CoT tokens or simply replicate the final response at the token level. In autonomous driving, generating even a shortened CoT trajectory incurs unacceptable latency, while standard token-level imitation fails to transfer the intricate spatial and logical reasoning deeply embedded in the teacher's cognitive process. To address this, we propose a new Latent Thought Distillation approach that directly aligns the student's hidden representations with the teacher's fully converged reasoning states, achieving an optimal balance between reasoning depth and inference efficiency.

3 Method

In this section, we describe the proposed CritiqueDriveVLM framework, designed to tackle two critical bottlenecks in autonomous driving VLMs: the inherent unreliability of reasoning in complex scenarios, and the unacceptable inference latency of explicit CoT generation. As illustrated in Figure 2, our framework is implemented through a three-stage training pipeline: **Stage 1** (Warm-up SFT and Verifier Construction) establishes the foundational reasoning format and builds a multi-dimensional verifier; **Stage 2** (Critique-Driven Multi-Turn RL) leverages verifier feedback and a multi-turn penalty to cultivate a highly reliable Teacher model capable of verifier-guided refinement; Finally, **Stage 3** (Latent Thought Distillation) transfers the Teacher's deep logical representations into a Student model, eliminating the latency of explicit CoT generation.

3.1 Warm-up SFT and Verifier Construction

Before initiating large-scale RL, it is essential to equip the base policy with a standardized CoT paradigm and construct an independent multi-dimensional verifier.

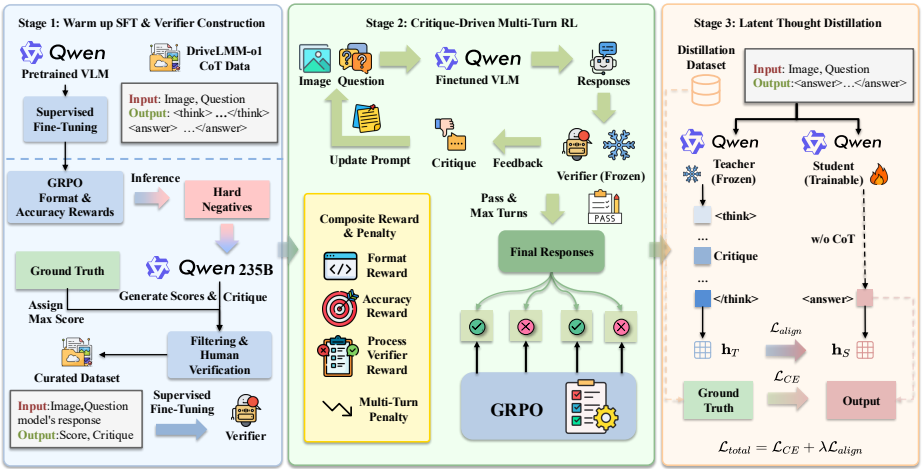

Fig. 2: Overall pipeline of the proposed CritiqueDriveVLM framework. (i) Stage 1: Warm-up SFT and Verifier Construction establishing a structural reasoning format and training a multi-dimensional verifier; (ii) Stage 2: Critique-Driven Multi-Turn RL cultivating a highly reliable Teacher via verifier feedback; and (iii) Stage 3: Latent Thought Distillation internalizing the Teacher's latent representations into a CoT-free, low-latency Student model.

Warm-up SFT. We perform a warm-up SFT on the base VLM using a subset of high-quality CoT data from the DriveLMM-o1 dataset [17]. The primary objective here is not to maximize the model's reasoning capabilities, but to strictly enforce a structural output format. Given an input image V and a user instruction x, the model is trained to consistently generate a complete response y formatted as $y =$ `<think>` r `</think>` `<answer>` a `</answer>`, where r denotes the intermediate reasoning process and a represents the final answer. This format initialization acts as a critical anchor, preventing mode collapse and format degradation when the model explores the expansive action space during the subsequent RL phase.

Multi-Dimensional Verifier Construction. The core component of our multi-turn framework is an independent, frozen multi-dimensional verifier $\mathcal{V}$, responsible for evaluating the model's complete generated response y. To train this verifier (initialized from a separate base model), we design a rigorous data generation pipeline. Positive samples are directly derived from ground truth annotations, which are inherently assigned maximum scalar scores across all dimensions with an empty natural language critique. To mine hard negatives, we deploy a baseline GRPO [49] model, trained solely with formatting and accuracy rewards, to perform inference on the training set. This yields a wealth of erroneous outputs exhibiting visual hallucinations or logical contradictions. These flawed responses are subsequently evaluated by a substantially more capable model

(Qwen3-VL-235B [1]), which assigns discrete multi-dimensional scores as follows:

$$R_{\text{verif}} = \{s_{\text{per}}, s_{\text{log}}, s_{\text{saf}}\}, \quad s \in \{0, 0.5, 1.0\}, \tag{1}$$

covering perception, logic, and safety. Concurrently, the Qwen3-VL-235B model generates targeted text-based critiques c addressing the specific errors. Following a stringent data filtering and human verification process, this curated dataset is used to train the verifier. During the RL stage, the function $\mathcal{V}(V, x, y) \to (R_{\text{verif}}, c)$ provides the policy model with granular scalar feedback and, if the response fails to meet requirements, it also generates a text-based critique prompt to guide the subsequent reasoning turn.

3.2 Critique-Driven Multi-Turn RL

To train a highly reliable Teacher model π_θ capable of profound logical reasoning and verifier-guided refinement, we introduce a multi-turn RL paradigm. Unlike traditional RL settings that rely solely on binary task-completion rewards, our framework leverages the frozen multi-dimensional verifier $\mathcal{V}$ (constructed in Stage 1) to provide fine-grained textual and scalar feedback. The definitions of the composite reward components are detailed in Table 1.

Multi-Turn Interaction. We formulate the reasoning process as a sequential interactive loop with a maximum allowed turn limit K. At the initial turn $t = 1$, given an input image V and a prompt instruction $x_1 = x$, the policy model generates a complete response y_1. The verifier $\mathcal{V}$ then evaluates y_1 to output a score set R_{verif} and a text critique c_1. If R_{verif} achieves a perfect score across all dimensions, or the interaction reaches the maximum limit ($t = K$), the loop terminates at turn T. Otherwise, the verifier outputs a critique c_t, which is appended to the dialogue history to construct a new user prompt: $x_{t+1} = \text{Concat}(x_t, y_t, c_t)$. The model then generates a refined response y_{t+1} based on the critique history [31, 51].

Composite Reward and Efficiency Penalty. The reward is calculated based on the terminal response y_T at the concluding turn T. As outlined in Table 1, the base reward R_{base} aggregates formatting, accuracy, and process-aware evaluations as follows:

$$R_{\text{base}}(y_T) = R_{\text{fmt}}(y_T) + R_{\text{acc}}(y_T) + \alpha(s_{\text{per}} + s_{\text{log}} + s_{\text{saf}}), \tag{2}$$

where α is a weighting coefficient for the verifier scores.

While multi-turn interaction significantly enhances reasoning accuracy, autonomous driving systems demand extremely low latency. To prevent the model from becoming overly reliant on external critiques, we introduce a step-decay multi-turn penalty $\mathcal{P}_{\text{mt}}(T)$. This penalty explicitly forces the model to internalize

Table 1: Definitions of the composite reward components. By isolating the reasoning process (`<think>`) and final decision (`<answer>`), our Multi-Turn RLVR provides fine-grained feedback to mitigate conservative bias and visual hallucinations.

Reward Type	**Description**
Format Reward (R_{fmt})	**Format Compliance:** Ensures the model's output strictly adheres to the predefined XML-style structure (*i.e.*, isolating reasoning within `<think>` tags and decisions within `<answer>` tags). Promotes reliable trajectory parsing.
Accuracy Reward (R_{acc})	**Final Answer Alignment:** Verifies the extracted final action directive (parsed from the `<answer>` block) against the ground truth. Promotes task-level decision correctness.
Process Verifier Reward (R_{verif})	Evaluates the intermediate CoT (parsed from the `<think>` block) through multi-dimensional scalar feedback. Sub-rewards for: **(i) Perception Score ($s_{\mathbf{per}}$):** Verifies visual entity grounding accuracy and penalizes visual hallucinations. **(ii) Logic Score ($s_{\mathbf{log}}$):** Evaluates the rigor of causal reasoning and penalizes internal logical contradictions. **(iii) Safety Score ($s_{\mathbf{saf}}$):** Assesses reasoning safety to mitigate conservative bias.
Multi-Turn Penalty ($\mathcal{P}_{\text{mt}}$)	**Efficiency Constraint:** A step-decay penalty applied based on the number of critique attempt turns. Forces the model to internalize reasoning and promotes first-attempt correctness.

its reasoning process and promotes first-attempt correctness. The final terminal reward is formulated as:

$$R_{\text{final}} = R_{\text{base}}(y_T) - \mathcal{P}_{\text{mt}}(T). \tag{3}$$

In practical implementation, to optimally balance exploration efficiency and computational overhead, we set the maximum turn limit to $K = 2$. Thus, the model is penalized with a constant decay value if it requires a second turn to correct its errors.

GRPO Optimization Objective. To avoid the massive memory overhead associated with maintaining an equivalent-sized Critic network in the standard Proximal Policy Optimization (PPO) algorithm [46], we employ GRPO in this work. For a given instruction x drawn from the training distribution $P(x)$, we sample a group of G terminal responses, denoted as $\{y_1, y_2, \ldots, y_G\}$, from the old policy $\pi_{\theta_{\text{old}}}$. The policy is optimized by maximizing the following GRPO

objective:

$$\mathcal{J}_{\text{GRPO}}(\theta) = \mathbb{E}_{x\sim P(x),\{y_i\}_{i=1}^{G}\sim\pi_{\theta_{\text{old}}}} \left[\frac{1}{G}\sum_{i=1}^{G} \mathcal{L}_i - \beta \mathbb{D}_{\text{KL}}(\pi_\theta \| \pi_{\text{ref}}) \right], \tag{4}$$

where the group-wise clipped loss $\mathcal{L}_i$ is defined as follows:

$$\mathcal{L}_i = \min\left(\rho_i A_i, \text{clip}(\rho_i, 1-\epsilon, 1+\epsilon) A_i\right), \tag{5}$$

where the importance sampling ratio ρ_i and the normalized advantage A_i are given by:

$$\rho_i = \frac{\pi_\theta(y_i|x)}{\pi_{\theta_{\text{old}}}(y_i|x)}, \tag{6}$$

$$A_i = \frac{R_{\text{final}}^{(i)} - \text{mean}(R_{\text{final}})}{\text{std}(R_{\text{final}})}. \tag{7}$$

The hyperparameter ϵ clips this ratio to ensure training stability. Additionally, β controls the strength of the Kullback-Leibler (KL) divergence penalty $\mathbb{D}_{\text{KL}}$, which restricts the policy π_θ from deviating excessively from the reference model π_{ref} established during the warm-up SFT stage.

3.3 Latent Thought Distillation

To resolve the prohibitive latency of explicit CoT generation, we propose Latent Thought Distillation, internalizing the Teacher's slow System-2 reasoning into the Student's fast System-1 intuition without outputting CoT text, inspired by [22, 61].

We initialize the Student model directly from the original base VLM. Furthermore, we construct the distillation dataset by strictly formatting the DriveLMM-o1 training data into pure prompt-answer pairs, entirely removing any explicit CoT text. This setup forces the Student to rely solely on its internal latent representations to maintain logical depth.

Our strategy aligns the hidden representation of the Student with the final reasoning state of the Teacher. Let π_T and π_S be the frozen Teacher and trainable Student, respectively. For a given input, π_T generates its multi-turn reasoning trajectory. We extract the deep hidden state at the last `</think>` token of the trajectory, denoted as $\mathbf{h}_T^{\text{think}} \in \mathbb{R}^d$. Simultaneously, π_S skips the intermediate reasoning steps to directly predict the final output. We extract its corresponding hidden state at the `<answer>` token, denoted as $\mathbf{h}_S^{\text{answer}} \in \mathbb{R}^d$.

To ensure the Student's internal logic mimics the Teacher's thought process, we employ Cosine Similarity as the alignment objective:

$$\mathcal{L}_{\text{align}} = 1 - \frac{\mathbf{h}_S^{\text{answer}} \cdot \mathbf{h}_T^{\text{think}}}{\|\mathbf{h}_S^{\text{answer}}\|_2 \|\mathbf{h}_T^{\text{think}}\|_2}. \tag{8}$$

This objective strictly focuses on aligning the semantic direction of the reasoning logic in high-dimensional space. The Student model is subsequently optimized through a joint loss function:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{CE}} + \lambda \mathcal{L}_{\text{align}}, \tag{9}$$

where $\mathcal{L}_{\text{CE}}$ is the standard auto-regressive cross-entropy loss for supervised answer prediction, and λ is the balancing coefficient. Through this unified objective, the Student inherits the Teacher's logical rigor within its latent space, achieving efficient and accurate prediction.

4 Experiments

In this section, we comprehensively evaluate the effectiveness of the proposed CritiqueDriveVLM framework in complex autonomous driving scenarios. The remainder of this section is organized as follows: We first detail the experimental setup in Section 4.1. Then, we present the main quantitative results on the DriveLMM-o1 benchmark [17], highlighting the state-of-the-art reasoning capabilities of our Stage-2 Teacher model in Section 4.2. Next, we analyze the inference efficiency and the impact of Latent Thought Distillation under CoT-free constraints in Section 4.3. We further conduct ablation studies to validate the critique-driven multi-turn reinforcement learning mechanisms in Section 4.4. Finally, we provide a qualitative analysis of our model's decision-making process in Section 4.5.

4.1 Experimental Setup

Datasets. To ensure a fair comparison and strictly validate the effectiveness of our framework, we exclusively utilize the DriveLMM-o1 [17] benchmark, which is built upon the foundational nuScenes [3] dataset. Specifically, we employ its training split comprising over 18,000 VQA pairs, where each sample provides step-by-step reasoning for complex driving scenarios. This unified dataset is consistently applied across all three stages of our framework: Warm-up SFT, Critique-Driven RL, and Latent Thought Distillation.

Evaluation Metrics. We adopt the official evaluation protocol of the DriveLMM-o1 benchmark, reporting the Overall Reasoning score and Multiple Choice Quality (MCQ) as primary metrics. Additionally, we evaluate fine-grained task-specific driving and scene detail metrics (*e.g.*, Risk Assessment, Traffic Rule Adherence, Scene Awareness, Relevance, and Missing Details). Further metric calculation details are provided in Appendix A.

Implementation Details. We employ Qwen3-VL-8B [1] as our base model for both the Teacher and the Student, keeping the vision encoder frozen across

Table 2: Main quantitative comparison on the DriveLMM-o1 benchmark. **Bold** indicates best. Underline indicates second best. Our Stage-2 Teacher achieves state-of-the-art results without external tool-use.

Vision Language Models	Driving Metrics (%) ↑			Scene Detail (%) ↑		Overall (%) ↑	
	Risk Assess.	Rule Adh.	Scene Aware.	Relevance	Missing	Reason.	MCQ
GPT-4o [18]	71.32	80.72	72.96	76.65	71.43	72.52	57.84
Ovis1.5-Gemma2-9B [29]	51.34	66.36	54.74	55.72	55.74	55.62	48.85
Mulberry-7B [59]	51.89	63.66	56.68	57.27	57.45	57.65	52.86
LLaVA-CoT [58]	57.62	69.01	60.84	62.72	60.67	61.41	49.27
LlamaV-O1 [53]	60.20	73.52	62.67	64.66	63.41	63.13	50.02
InternVL2.5-8B [4]	69.02	78.43	71.52	75.80	70.54	71.62	54.87
Qwen2.5-VL-7B [2]	46.44	60.45	51.02	50.15	52.19	51.77	37.81
Qwen2.5-VL-72B [2]	64.40	72.81	60.29	65.13	62.81	65.73	61.27
Qwen3-VL-8B [1]	79.50	84.32	80.41	84.43	75.32	77.76	55.54
DriveLMM-o1 [17]	73.01	81.56	75.39	79.42	74.49	75.24	62.36
AgentThink [42]	80.51	84.98	82.11	<u>84.99</u>	**79.56**	79.68	71.35
OmniDrive-R1 [62]	**82.31**	<u>85.42</u>	**83.75**	82.58	78.26	<u>80.35</u>	<u>73.62</u>
Ours-Teacher(Stage 2)	<u>81.63</u>	**86.21**	<u>82.27</u>	**85.46**	<u>79.17</u>	**80.48**	**76.54**

all training phases. During Stage 1, the Warm-up SFT is conducted using Low-Rank Adaptation (LoRA) [15] with a learning rate of 1×10^{-4}. For Stage 2, the GRPO [49] fine-tuning utilizes a learning rate of 2×10^{-6} with 4 rollouts per question ($G = 4$), and the maximum multi-turn interaction limit is set to $K = 2$. In Stage 3, the Latent Thought Distillation is also optimized using LoRA with a learning rate of 2×10^{-5}, and the balancing coefficient for the alignment loss is set to $\lambda = 0.5$. Across all three training phases, the models are optimized for 2 epochs with a global batch size of 128. All experiments are conducted on a single server equipped with 8 NVIDIA A100 GPUs, and inference time is tested on a single A100 GPU. Additional implementation details and hyperparameter settings are elaborated in Appendix B.

4.2 Main Results

We evaluate the comprehensive driving and reasoning capabilities of our framework against a wide range of state-of-the-art VLMs on the DriveLMM-o1 benchmark. As shown in Table 2, the comparative baselines include prominent general-purpose models (*e.g.*, GPT-4o [18], InternVL2.5-8B [4], and the Qwen-VL series [1,2]) as well as domain-specific autonomous driving models (*e.g.*, DriveLMM-o1 [17], AgentThink [42], and OmniDrive-R1 [62]).

Our Stage-2 Teacher establishes a new state-of-the-art across multiple critical metrics. Notably, it achieves the highest Overall Reasoning score of 80.48% and an MCQ of 76.54%, significantly outperforming strong driving-specific counterparts like OmniDrive-R1 (73.62%) and AgentThink (71.35%). Unlike baselines relying

Table 3: Comparison of inference efficiency and ablation on Latent Thought Distillation (Stage 3). We evaluate the trade-off between reasoning accuracy and latency. The ablation of the alignment objective ($\mathcal{L}_{\text{align}}$) demonstrates that our Student model preserves reasoning depth while achieving low-latency execution under CoT-free constraints.

Models	CoT	Tool	$\mathcal{L}_{\text{align}}$	Avg. Tokens ↓	Avg. Time (ms) ↓	MCQ (%) ↑
System-2 Models (with CoT)						
Qwen3-VL-8B	✓	×	×	390.95	4265	55.54
Qwen3-VL-8B (SFT)	✓	×	×	147.88	1675	62.86
DriveLMM-o1	✓	×	×	150.52	1968	62.36
AgentThink	✓	✓	×	515.01	5416	71.35
Ours-Teacher(Stage 2)	✓	×	×	223.32	3482	**76.54**
System-1 Models (CoT-Free)						
Qwen3-VL-8B	×	×	×	37.23	461	54.96
Qwen3-VL-8B (SFT)	×	×	×	23.65	343	61.73
DriveLMM-o1	×	×	×	30.96	425	60.14
Ours-Student(Stage 3)	×	×	✓	28.83	416	**68.59**

on complex external APIs or interleaved multi-modal CoT, our Teacher achieves superior accuracy entirely through internalized logic cultivated during the critique-driven RL phase, avoiding explicit tool-use overhead.

Delving into the fine-grained metrics, our model demonstrates a significant advantage in Traffic Rule Adherence (86.21%) and Relevance (85.46%). While OmniDrive-R1 marginally leads in Risk Assessment (82.31%) and Scene Awareness and Object Understanding (83.75%), and AgentThink slightly edges out in the Missing Details metric (79.56%), our Teacher model maintains highly competitive and robust scores in these areas (81.63%, 82.27%, and 79.17%, respectively). This comprehensive and well-rounded performance profile proves that providing granular, multi-dimensional verifier feedback during RL effectively suppresses visual hallucinations and conservative biases, resulting in highly reliable end-to-end driving decisions.

4.3 Efficiency and Latent Thought Distillation

Real-time autonomous driving systems impose stringent constraints on inference latency. While explicit CoT generation and tool-use mechanisms significantly enhance reasoning accuracy, they inherently incur unacceptable computational overhead. As shown in the System-2 section of Table 3, traditional reasoning models suffer from extreme verbosity. AgentThink incurs a staggering latency of 5416 ms due to explicit tool invocations. Even our highly capable Stage-2 Teacher requires 3482 ms to articulate its internal reasoning process. To eliminate this latency bottleneck, we must evaluate models under a strict CoT-free constraint.

Table 4: Ablation study on Critique-Driven Multi-Turn RL (Stage 2). We progressively evaluate the impact of Accuracy Rewards (Acc), Multi-dimensional Verifier Rewards (Ver), and Multi-Turn Interaction & Penalty (MT).

Model Variant	SFT	Reward Setting			Driving Metrics (%) ↑			Detail (%) ↑		Overall (%) ↑	
		Acc	Ver	MT	Risk Assess.	Rule Adh.	Scene Aware.	Relev.	Miss.	Reason.	MCQ
Base Model	×	×	×	×	79.50	84.32	80.41	84.43	75.32	77.76	55.54
+ SFT	✓	×	×	×	73.74	79.46	75.19	79.70	71.24	73.38	62.86
+ GRPO	✓	✓	×	×	74.14	80.12	75.59	80.21	71.63	73.85	64.49
+ GRPO†	✓	✓	✓	×	75.30	81.20	76.84	81.23	72.89	75.10	67.46
Ours-Teacher	✓	✓	✓	✓	**81.63**	**86.21**	**82.27**	**85.46**	**79.17**	**80.48**	**76.54**

However, when standard baselines are forced to predict final answers directly without CoT, their performance drops precipitously. As shown in the System-1 section of Table 3, the Qwen3-VL-8B (SFT) model without distillation severely truncates the logical depth, achieving an MCQ of only 61.73%. To bridge this capability gap, our Latent Thought Distillation integrates an alignment objective ($\mathcal{L}_{\text{align}}$). By explicitly aligning the Student's latent representations with the Teacher's fully converged reasoning state, our Stage-3 Student achieves a superior MCQ of 68.59% using an average of only 28.83 tokens. This drastically slashes the absolute inference time to 416 ms, an 88% reduction compared to the Stage-2 Teacher (3482 ms), demonstrating that our proposed distillation method successfully internalizes deep System-2 reasoning capabilities into the fast System-1 latent space, achieving an optimal balance between reasoning accuracy and low-latency execution.

4.4 Ablation on Critique-Driven Multi-Turn RL

To validate the contribution of each core mechanism in Critique-Driven Multi-Turn RL (Stage 2), we ablate our reward design in Table 4. Starting from the Warm-up SFT baseline (62.86% MCQ), applying standard single-turn GRPO with only Accuracy Rewards yields a marginal improvement (64.49% MCQ). The addition of Multi-dimensional Verifier Rewards further boosts the MCQ to 67.46% while steadily improving fine-grained metrics such as Risk Assessment and Traffic Rule Adherence, indicating that providing granular scalar feedback effectively suppresses visual hallucinations. Crucially, introducing the Multi-Turn Interaction combined with a step-decay penalty significantly elevates the overall MCQ to 76.54%. This substantial improvement confirms our hypothesis: transitioning from a single-turn setup to a multi-turn loop allows the model to iteratively refine its reasoning based on verifier critiques, while the multi-turn penalty prevents over-reliance on these external cues, forcing the policy to internalize the logical deduction process for correctness.

4.5 Qualitative Analysis

To intuitively demonstrate the efficacy of our CritiqueDriveVLM framework, Figure 3 shows a comparative case study in a complex driving scenario involving

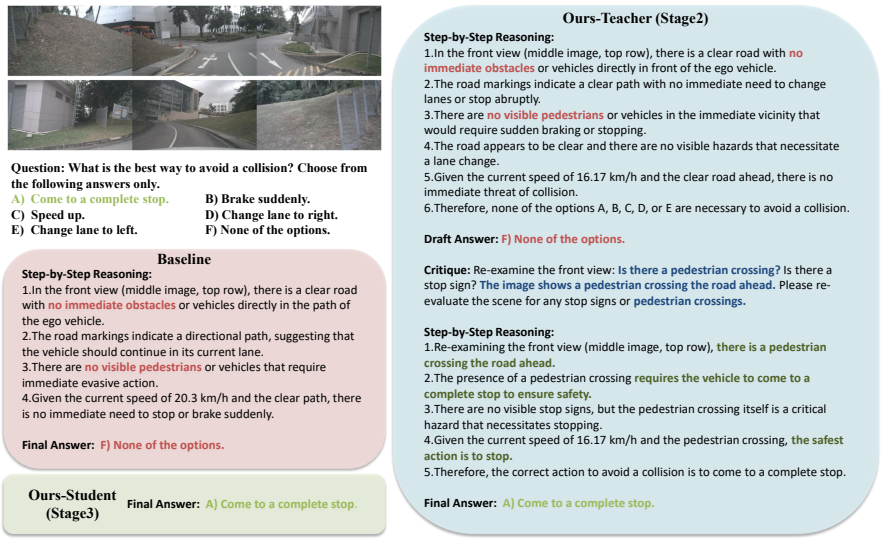

Fig. 3: Qualitative comparison in a pedestrian crossing scenario. The **Baseline** hallucinates and misses the pedestrian. Our **Stage-2 Teacher** corrects this error through verifier-guided critiques, while our **Stage-3 Student** directly predicts the safe action without explicit CoT overhead.

a pedestrian crossing. As illustrated, the standard SFT Baseline suffers from severe visual hallucination; it explicitly reasons that there are "no visible pedestrians" and dangerously concludes that no evasive action is required (Option F). In contrast, our Stage-2 Teacher overcomes this failure through its critique-driven multi-turn reasoning mechanism. While its initial draft mirrors the Baseline's dangerous oversight, the external verifier injects targeted critique feedback, alerting the model to the overlooked pedestrian. This prompts the Teacher to successfully re-examine the visual evidence, correct its flawed trajectory, and output the safe action to "Come to a complete stop" (Option A). Crucially, our Stage-3 Student directly and instantly predicts the correct, safe action (Option A) without any explicit reasoning overhead. This confirms that our Latent Thought Distillation effectively compresses the Teacher's rigorous System-2 risk assessment capabilities directly into the Student's fast System-1 latent space.

5 Conclusion

In this paper, we presented **CritiqueDriveVLM**, a unified three-stage framework designed to resolve the reasoning fragility and inference latency bottlenecks in autonomous driving VLMs. By introducing Critique-Driven Multi-Turn RL, our Teacher model internalized logical deduction to mitigate visual hallucinations, achieving state-of-the-art accuracy without relying on fragile external tools. Furthermore, our Latent Thought Distillation successfully compressed these deep

System-2 reasoning capabilities into a CoT-free System-1 Student, preserving decision accuracy while reducing inference latency to low-latency levels (416 ms). In the future, we plan to extend this framework to multi-camera video streams for temporal logic distillation.

Acknowledgements

This work is partly supported by the Funds for the Beijing Natural Science Foundation under Grant L243027, the NSFC Project under Grant 62572072 and Hainan Provincial Natural Science Foundation.

References

1. Bai, S., Cai, Y., Chen, R., et al.: Qwen3-VL technical report. arXiv preprint arXiv:2511.21631 (2025)
2. Bai, S., Chen, K., Liu, X., et al.: Qwen2.5-VL technical report. arXiv preprint arXiv:2502.13923 (2025)
3. Caesar, H., Bankiti, V., Lang, A.H., Vora, S., Liong, V.E., Xu, Q., Krishnan, A., Pan, Y., Baldan, G., Beijbom, O.: nuScenes: A multimodal dataset for autonomous driving. In: IEEE Conf. Comput. Vis. Pattern Recog. (2020)
4. Chen, Z., Wang, W., Cao, Y., et al.: Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. arXiv preprint arXiv:2412.05271 (2025)
5. Deng, W., Zhang, X., Qi, M.: Active exploring like a pigeon: Reinforcing spatial reasoning via agentic vision-language models. In: Int. Conf. Mach. Learn. (2026)
6. Deng, Y., Prasad, K., Fernandez, R., et al.: Implicit chain of thought reasoning via knowledge distillation. arXiv preprint arXiv:2311.01460 (2023)
7. Ettinger, S., Cheng, S., Caine, B., Liu, C., Zhao, H., Pradhan, S., Chai, Y., Sapp, B., Qi, C.R., Zhou, Y., et al.: Large scale interactive motion forecasting for autonomous driving: The Waymo open motion dataset. In: Int. Conf. Comput. Vis. pp. 9710–9719 (2021)
8. Fan, H., Zhu, F., Liu, C., Zhang, L., et al.: Baidu Apollo EM motion planner. arXiv preprint arXiv:1807.08048 (2018)
9. Gao, J., Sun, C., Zhao, H., et al.: VectorNet: Encoding HD maps and agent dynamics from vectorized representation. In: IEEE Conf. Comput. Vis. Pattern Recog. (2020)
10. Gu, Y., Dong, L., Wei, F., Huang, M.: MiniLLM: Knowledge distillation of large language models. In: Int. Conf. Learn. Represent. (2024)
11. Guo, D., Yang, D., Zhang, H., et al.: DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. arXiv preprint arXiv:2501.12948 (2025)
12. Hao, S., Sukhbaatar, S., Su, D., et al.: Training large language models to reason in a continuous latent space. arXiv preprint arXiv:2412.06769 (2024)
13. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531 (2015)
14. Hsieh, C.Y., Li, C.L., Yeh, C.K., et al.: Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes. In: Findings Assoc. Comput. Linguist. (ACL) (2023)
15. Hu, E.J., Shen, Y., Wallis, P., et al.: LoRA: Low-rank adaptation of large language models. In: Int. Conf. Learn. Represent. (2022)

16. Hu, Y., Yang, J., Chen, L., et al.: Planning-oriented autonomous driving. In: IEEE Conf. Comput. Vis. Pattern Recog. (2023)
17. Ishaq, A., Lahoud, J., More, K., et al.: DriveLMM-o1: A step-by-step reasoning dataset and large multimodal model for driving scenario understanding. In: IEEE/RSJ Int. Conf. Intell. Robots Syst. (2025)
18. Islam, R., Moushi, O.M.: GPT-4o: The cutting-edge advancement in multimodal LLM. In: Proc. Comput. Conf. pp. 47–60. Springer (2025)
19. Jiang, B., Chen, S., Xu, Q., et al.: VAD: Vectorized scene representation for efficient autonomous driving. In: Int. Conf. Comput. Vis. (2023)
20. Jiang, B., Chen, S., Zhang, Q., Liu, W., Wang, X.: AlphaDrive: Unleashing the power of VLMs in autonomous driving via reinforcement learning and reasoning. arXiv preprint arXiv:2503.07608 (2025)
21. Jiang, Y., Xiong, Y., Yuan, Y., et al.: PAG: Multi-turn reinforced LLM self-correction with policy as generative verifier. arXiv preprint arXiv:2506.10406 (2025)
22. Kahneman, D.: Thinking, Fast and Slow. Macmillan (2011)
23. Levinson, J., Askeland, J., Becker, J., et al.: Towards fully autonomous driving: Systems and algorithms. In: Proc. IEEE Intell. Veh. Symp. (IV) (2011)
24. Li, Y., Du, Y., Zhou, K., et al.: Evaluating object hallucination in large vision-language models. In: Conf. Empir. Methods Nat. Lang. Process. (2023)
25. Li, Z., Wang, W., Li, H., et al.: BEVFormer: Learning Bird's-Eye-View representation from multi-camera images via spatiotemporal transformers. In: Eur. Conf. Comput. Vis. (2022)
26. Liao, D., Qi, M., Liu, L., Ma, H.: Improving batch normalization with test-time adaptation for robust object detection in self-driving. In: AAAI (2026)
27. Lightman, H., Kosaraju, V., Burda, Y., et al.: Let's verify step by step. In: Int. Conf. Learn. Represent. (2024)
28. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. In: Adv. Neural Inform. Process. Syst. (2023)
29. Lu, S., Li, Y., Chen, Q.G., et al.: Ovis: Structural embedding alignment for multi-modal large language model. arXiv preprint arXiv:2405.20797 (2024)
30. Lv, C., Qi, M., Liu, L., Ma, H.: T2SG: Traffic topology scene graph for topology reasoning in autonomous driving. In: IEEE Conf. Comput. Vis. Pattern Recog. (2025)
31. Madaan, A., Tandon, N., Gupta, P., et al.: Self-Refine: Iterative refinement with self-feedback. In: Adv. Neural Inform. Process. Syst. (2023)
32. Nie, M., Peng, R., Wang, C., et al.: Reason2Drive: Towards interpretable and chain-based reasoning for autonomous driving. In: Eur. Conf. Comput. Vis. (2024)
33. Ouyang, L., Wu, J., Jiang, X., et al.: Training language models to follow instructions with human feedback. In: Adv. Neural Inform. Process. Syst. (2022)
34. Qi, M., Lv, C., Ma, H.: Robust disentangled counterfactual learning for physical audiovisual commonsense reasoning. IEEE Trans. Pattern Anal. Mach. Intell. (2026)
35. Qi, M., Peng, J., Zhang, X., Ma, H.: Towards balanced multi-modal learning in 3D human pose estimation. In: IEEE Conf. Comput. Vis. Pattern Recog. (2026)
36. Qi, M., Qin, J., Yang, Y., Wang, Y., Luo, J.: Semantics-aware spatial-temporal binaries for cross-modal video retrieval. IEEE Trans. Image Process. (2021)
37. Qi, M., Qin, J., Zhen, X., Huang, D., Yang, Y., Luo, J.: Few-shot ensemble learning for video classification with SlowFast memory networks. In: ACM Int. Conf. Multimedia (2020)
38. Qi, M., Wang, Y., Li, A., Luo, J.: Sports video captioning via attentive motion representation and group relationship modeling. IEEE Trans. Circuit Syst. Video Technol. (2019)

39. Qi, M., Wu, Y., Yun, W., Zhang, X., Ma, H.: Explainable action form assessment by exploiting multimodal chain-of-thoughts reasoning. IEEE Trans. Image Process. (2026)
40. Qi, M., Ye, H., Peng, J., Ma, H.: Action quality assessment via hierarchical pose-guided multi-stage contrastive regression. IEEE Trans. Image Process. (2025)
41. Qi, M., Zhu, P., Li, X., Bi, X., Qi, L., Ma, H., Yang, M.H.: DC-SAM: In-context segment anything in images and videos via dual consistency. IEEE Trans. Pattern Anal. Mach. Intell. (2026)
42. Qian, K., Jiang, S., Zhong, Y., et al.: AgentThink: A unified framework for tool-augmented chain-of-thought reasoning in vision-language models for autonomous driving. In: Findings Assoc. Comput. Linguist. (EMNLP) (2025)
43. Radford, A., Kim, J.W., Hallacy, C., et al.: Learning transferable visual models from natural language supervision. In: Int. Conf. Mach. Learn. (2021)
44. Romero, A., Ballas, N., Kahou, S.E., et al.: FitNets: Hints for thin deep nets. In: Int. Conf. Learn. Represent. (2015)
45. Rowe, L., de Schaetzen, R., Girgis, R., et al.: Poutine: Vision-language-trajectory pre-training and reinforcement learning post-training enable robust end-to-end autonomous driving. arXiv preprint arXiv:2506.11234 (2025)
46. Schulman, J., Wolski, F., Dhariwal, P., Radford, A., Klimov, O.: Proximal policy optimization algorithms. arXiv preprint arXiv:1707.06347 (2017)
47. Shao, H., Hu, Y., Wang, L., et al.: LMDrive: Closed-loop end-to-end driving with large language models. In: IEEE Conf. Comput. Vis. Pattern Recog. (2024)
48. Shao, Z., Luo, Y., Lu, C., et al.: DeepSeekMath-v2: Towards self-verifiable mathematical reasoning. arXiv preprint arXiv:2511.22570 (2025)
49. Shao, Z., Wang, P., Zhu, Q., et al.: DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. arXiv preprint arXiv:2402.03300 (2024)
50. Shen, Z., Yan, H., Zhang, L., et al.: CODI: Compressing chain-of-thought into continuous space via self-distillation. In: Conf. Empir. Methods Nat. Lang. Process. (2025)
51. Shinn, N., Cassano, F., Gopinath, A., Narasimhan, K., Yao, S.: Reflexion: Language agents with verbal reinforcement learning. In: Adv. Neural Inform. Process. Syst. (2023)
52. Sima, C., Renz, K., Chitta, K., et al.: DriveLM: Driving with graph visual question answering. In: Eur. Conf. Comput. Vis. (2024)
53. Thawakar, O., Dissanayake, D., More, K.P., et al.: LlamaV-o1: Rethinking step-by-step visual reasoning in LLMs. In: Findings Assoc. Comput. Linguist. (ACL) (2025)
54. Tian, X., Gu, J., Li, B., Liu, Y., Zhao, Z., Wang, Y., Zhan, K., Jia, P., Lang, X., Zhao, H.: DriveVLM: The convergence of autonomous driving and large vision-language models. arXiv preprint arXiv:2402.12289 (2024)
55. Wang, W., Xie, J., Hu, C., et al.: DriveMLM: Aligning multi-modal large language models with behavioral planning states for autonomous driving. arXiv preprint arXiv:2312.09245 (2023)
56. Wang, Y., Kordi, Y., Mishra, S., et al.: Self-Instruct: Aligning language models with self-generated instructions. In: Ann. Meet. Assoc. Comput. Linguist. (2023)
57. Wei, J., Wang, X., Schuurmans, D., et al.: Chain-of-thought prompting elicits reasoning in large language models. In: Adv. Neural Inform. Process. Syst. (2022)
58. Xu, G., Jin, P., Wu, Z., et al.: LLaVA-CoT: Let vision language models reason step-by-step. In: Int. Conf. Comput. Vis. (2025)

59. Yao, H., Huang, J., Wu, W., et al.: Mulberry: Empowering MLLM with o1-like reasoning and reflection via collective Monte Carlo tree search. arXiv preprint arXiv:2412.18319 (2024)
60. Ye, H., Qi, M., Liu, Z., Liu, L., Ma, H.: SafeDriveRAG: Towards safe autonomous driving with knowledge graph-based retrieval-augmented generation. In: ACM Int. Conf. Multimedia (2025)
61. Yu, P., Xu, J., Weston, J., Kulikov, I.: Distilling System 2 into System 1. arXiv preprint arXiv:2407.06023 (2024)
62. Zhang, Z., Zheng, H., Wang, Y., et al.: OmniDrive-R1: Reinforcement-driven interleaved multi-modal chain-of-thought for trustworthy vision-language autonomous driving. arXiv preprint arXiv:2512.14044 (2025)
63. Zhou, Z., Cai, T., Zhao, S.Z., et al.: AutoVLA: A vision-language-action model for end-to-end autonomous driving with adaptive reasoning and reinforcement fine-tuning. arXiv preprint arXiv:2506.13757 (2025)
64. Zhu, P., Qi, M., Li, X., Li, W., Ma, H.: Unsupervised self-driving attention prediction via uncertainty mining and knowledge embedding. In: Int. Conf. Comput. Vis. (2023)