RSPO: Reasoning-Supervised Policy Optimization for Long-Tail Autonomous Driving

Jiayi Guan*[1,2] Jing Wu*[1] Jianhua Wu*[1,2] Jinghui Lu*[1] Zhijian Huang*[1,2] Guang Li [1]
Xiaoshuai Hao†[1] Hanbing Li [1] Yigu Ge [1] Tao Xu [1,2] Haiyang Sun [1] Bing Wang‡[1,2]
Guang Chen [1,2] Kuiyuan Yang [1,2] Hangjun Ye [1,2] Long Chen†[1]

[1] Xiaomi EV; [2] Beijing Key Laboratory of Intelligent Driving Foundation Models and Simulation

{guanjiayi, wujianhua1, huangzhijian5, xutao21, wangbing20, haoxiaoshuai, chenlong37}@xiaomi.com
edata18.ai@gmail.com, jinghui.lu2026@gmail.com

Abstract

Vision-Language Models (VLMs) have achieved remarkable advances in general autonomous driving. Nevertheless, effectively exploiting the reasoning capability of Chain-of-Thought (CoT) to enhance the robustness and accuracy of decision and planning in long-tail scenarios still remains a challenging open problem. To this end, we propose a **R**easoning-**S**upervised **P**olicy **O**ptimization (RSPO) algorithm to boost reasoning and decision-making performance under long-tail driving scenarios. Concretely, we first conduct an empirical analysis of the core challenges faced by current VLMs in autonomous driving decision and planning tasks, and formally define reasoning and decision-making tasks tailored to long-tail scenarios. On this basis, we design a curriculum-guided supervised fine-tuning strategy to enable the model to rapidly fit the mean distribution of model outputs. Furthermore, we propose a reasoning-supervised optimization algorithm, which enhances the robustness and accuracy of reasoning and decision in long-tail scenarios by guiding the optimality consistency between reasoning processes and final decisions. Finally, we construct multiple verifiable autonomous driving reasoning and decision datasets based on open-source benchmarks and conduct extensive multi-dimensional comparative experiments. Experimental results show that the proposed algorithm achieves consistent improvements in prediction accuracy across multiple datasets compared with existing methods. In particular, it achieves a 6.7% relative improvement over the strong baseline on the CODA-LM task and also yields lower trajectory-prediction errors on downstream planning tasks.

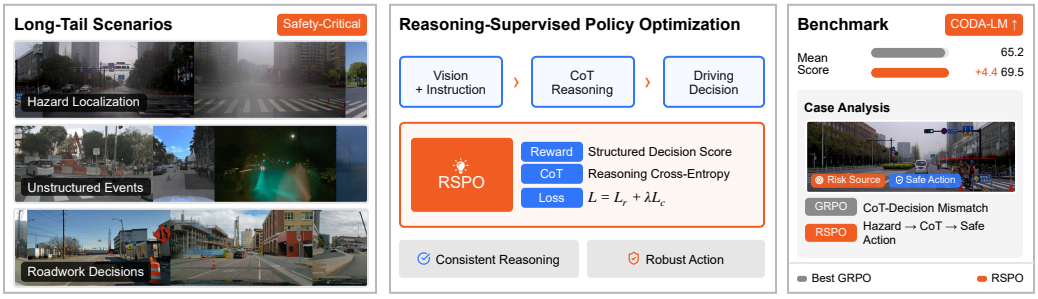


Figure 1: Teaser of RSPO for long-tail autonomous driving. RSPO combines chain-of-thought supervision with reward-guided policy optimization to align reasoning trajectories with driving decisions, yielding a clear CODA-LM mean-score improvement and more consistent hazard-to-action reasoning in representative long-tail cases.

*Equal contribution.
†Corresponding author.
‡Project Lead.

Conference on Neural Information Processing Systems (NeurIPS).

1 Introduction

Vision-Language Models (VLMs) have recently advanced embodied intelligence [5, 18, 19, 32, 59] and autonomous driving [17, 25, 27, 53, 64] through their strong knowledge-driven reasoning capabilities. In autonomous driving, they can integrate heterogeneous sensory inputs, perform contextual reasoning, and transfer knowledge across scenarios, showing strong potential in end-to-end vision-language-action (VLA) planning [10, 15, 23, 24, 35], scene understanding [20, 43, 45], and decision control [22, 33, 36]. However, existing VLA-based systems remain limited in long-tail scenarios, such as extreme weather and sudden obstacles, where low-probability but high-risk events pose major challenges to reliable deployment. Although chain-of-thought (CoT) [48] reasoning can integrate multi-source perceptual information and mine latent scene cues, existing methods still suffer from CoT-decision inconsistency and limited reasoning robustness. It remains challenging to jointly improve CoT quality and decision optimality without additional evaluation models, calling for targeted optimization strategies for long-tail autonomous driving reasoning and decision-making. To address this challenge, we propose Reasoning-Supervised Policy Optimization (RSPO), whose overall framework and core advantages are summarized in Fig. 1.

For CoT reasoning and decision tasks in long-tail autonomous driving scenarios, reinforcement fine-tuning (RFT) based on supervised fine-tuning (SFT) has exhibited substantial application potential [1, 11, 16, 21, 49]. However, when mainstream RFT algorithms such as GRPO [41] and DAPO [54] are applied to CoT-oriented tasks, they mainly optimize the final decision while disregarding the intermediate reasoning process. This outcome-centric objective may lead to overfitting to the training distribution and weaken long-tail reasoning robustness. Sequence-level strategies like GSPO [62] optimize the entire reasoning trajectory, but full-sequence optimization tends to increase predictive entropy and hinder convergence to optimal decisions. Recent works including GRMs [60] and R1-Reward [61] introduce an extra reward model, generally comparable in scale to the target multimodal model, to jointly evaluate CoT reasoning and final decisions. Nevertheless, their performance depends heavily on reward-model accuracy, and the additional model significantly increases training overhead.

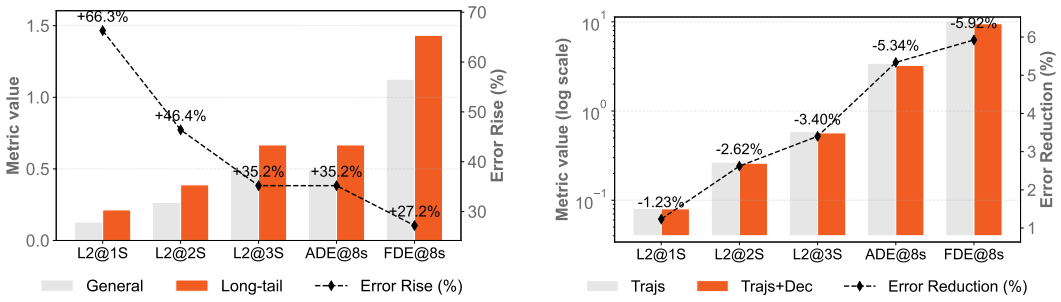
A fundamental challenge is to jointly improve the quality of CoT reasoning and the precision of decision in large multimodal models without introducing an additional reward evaluation model, which is essential for robust autonomous driving in long-tail scenarios. To this end, we propose a reasoning-supervised policy optimization algorithm for autonomous driving reasoning and decision. We first empirically analyze the core limitations of existing driving policies built upon VLMs and formally define the reasoning and decision tasks for long-tail scenarios. On this basis, we design a curriculum-guided supervised fine-tuning strategy to rapidly learn the mapping from long-tail scenarios to structured reasoning and decision outputs. Furthermore, we propose a reasoning-supervised policy optimization algorithm that improves reasoning and decision robustness by maximizing decision rewards while minimizing the cross-entropy loss of CoT reasoning. Finally, we construct verifiable reasoning and decision datasets based on multiple open-source autonomous driving datasets. Extensive multidimensional experimental comparisons demonstrate that in long-tail scenarios[4], our proposed method achieves steady accuracy gains in reasoning and decision-making over existing approaches, and also reduces trajectory-prediction errors in downstream planning tasks. The main contributions of this work are listed as follows:

- We empirically show that VLM-based driving policies suffer increased planning errors in long-tail scenarios, while explicit decision reasoning reduces trajectory-prediction errors. Based on these findings, we formalize long-tail autonomous driving reasoning and decision-making as a verifiable task and curate multiple datasets from open-source data.
- We propose a reasoning-supervised policy optimization algorithm for autonomous driving reasoning and decision. This algorithm maximizes decision rewards while minimizing the cross-entropy loss of CoT reasoning, promotes consistency between reasoning trajectories and decision outcomes without introducing an additional reward model, and further improves the robustness and accuracy of model reasoning and decision in long-tail scenarios.
- We conduct extensive comparative and ablation experiments on multiple validated long-tail datasets. Experimental results demonstrate steady improvements in decision accuracy across all datasets compared with existing methods, including a 6.7% relative improvement over the baseline on CODA-LM [6] and lower trajectory-prediction errors on downstream tasks.

[4]Our code is available at the following GitHub repository: github.RSPO.

2 Preliminary

Despite recent progress in autonomous driving, multimodal large language models still show limited reliability in long-tail scenarios. We conduct two preliminary planning studies to quantify this limitation and the role of reasoning. The first study evaluates policies trained on general data under both general and long-tail splits from the same dataset. As shown in Figure 2(a), these policies suffer substantial L2/ADE/FDE error increases on long-tail cases, indicating weak robustness under distributional shift. The second study compares trajectory-only prediction with joint decision-trajectory prediction on AR1 [47]; dataset construction and evaluation details are provided in Appendix A.1.1. Figure 2(b) shows that adding explicit decision reasoning reduces planning errors, suggesting that reasoning and decision outcomes aligned with the planning objective are beneficial for long-tail planning. Together, these results indicate that long-tail planning requires both robust visual generalization and decision-aware reasoning, rather than trajectory fitting alone.



(a) Impromptu dataset (b) AR1 dataset

Figure 2: Preliminary studies on planning performance. (a) Comparison between general and long-tail scenarios on the Impromptu dataset, revealing significant performance degradation in long-tail cases. (b) Comparison between trajectory-only output (*Trajs*) and joint decision-trajectory output (*Trajs+Dec*) on the AR1 dataset, demonstrating that incorporating explicit decision reasoning improves planning quality.

These observations motivate a task formulation that couples reasoning with decision-making for long-tail driving. While future trajectories are typically available in existing data, accurate reasoning and decision labels remain difficult to obtain, especially when causal scene understanding and driving actions must be aligned. We therefore define an autonomous driving reasoning and decision-making task to support automatic label construction and subsequent planning-model training.

Definition 1. *The long-tail autonomous driving reasoning and decision-making task takes as input one or multiple front-view images, the current vehicle state, and navigation information. It detects decision-relevant objects and predicts the future driving decision with its underlying rationale. Formally, it is represented as follows:*

$$[\mathcal{T}_{y_d}, \mathcal{T}_{y_c}] = \arg\max P_\theta(\mathcal{T}_{y_d}, \mathcal{T}_{y_c} | \mathcal{T}_v, \mathcal{T}_l), \tag{1}$$

where $\mathcal{T}_{y_c}$ denotes the predicted chain-of-thought reasoning tokens, including key object detection and the underlying rationale; $\mathcal{T}_{y_d}$ denotes the predicted decision-result tokens; $\mathcal{T}_l$ denotes the current vehicle state and navigation information; and $\mathcal{T}_v$ denotes one or multiple front-view images of the ego vehicle.

Definition 1 specifies the objective of reasoning and decision in long-tail safety-critical scenarios, providing an effective formulation for automatic label annotation and planning-model pretraining. By separating CoT tokens from decision tokens, it also matches the optimization target used in the following RSPO algorithm and turns the preliminary findings into a concrete supervision interface.

3 Methodology

Through the empirical analysis in Section 2, we have clarified the research objective of reasoning and decision in long-tail autonomous driving scenarios. To optimize the objective defined in Definition 1, we adopt a staged training paradigm inspired by previous work. Concretely, in the first stage, we

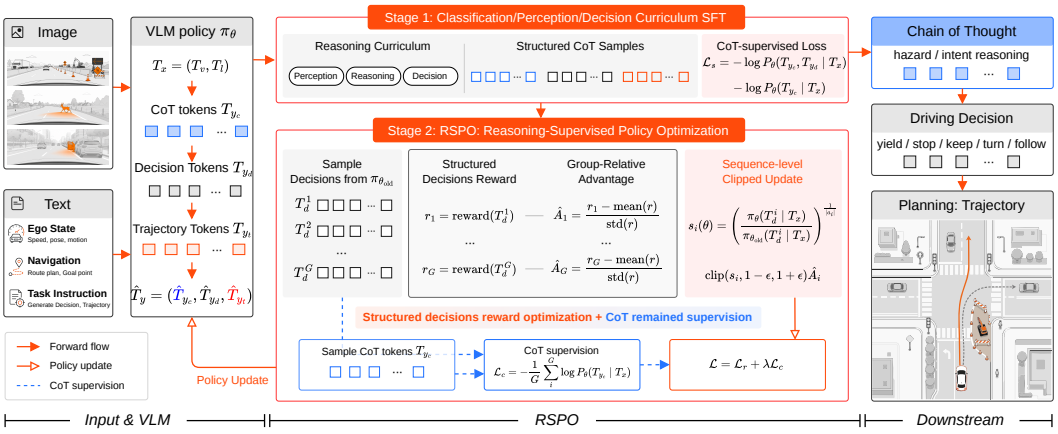


Figure 3: Overview of the proposed two-stage RSPO framework. The policy is first initialized by curriculum-guided SFT on structured CoT and decision labels, then optimized with rule-based decision rewards and CoT supervision. This joint objective improves reasoning-aligned driving decisions for long-tail autonomous driving scenarios.

design a Curriculum-Guided Supervised Fine-Tuning (CG-SFT) strategy to efficiently learn the mapping between long-tail scenarios and their corresponding reasoning and decision labels from sample data. In the second stage, we propose a novel Reasoning-Supervised Policy Optimization (RSPO) algorithm to further explore more robust reasoning and decision policies. The overall framework is illustrated in Figure 3. We detail the CG-SFT strategy in Section 3.1 and describe the RSPO algorithm in Section 3.2.

3.1 Curriculum-Guided Supervised Fine-Tuning Strategy

Following the optimization objective in Definition 1, this strategy employs supervised fine-tuning to efficiently learn the mapping from perceptual inputs, vehicle state, and navigation information to driving-oriented reasoning and decision-making. Given the complexity of long-horizon decision-making in autonomous driving and the intrinsic dependency between reasoning and decision outcomes, where decisions are conditioned on scene understanding and root-cause reasoning, we decompose the objective in Definition 1 into the following conditional probability formulation:

$$P(\mathcal{T}_{y_d}, \mathcal{T}_{y_c} | \mathcal{T}_v, \mathcal{T}_l) = P(\mathcal{T}_{y_d}, | \mathcal{T}_{y_c}, \mathcal{T}_v, \mathcal{T}_l) \cdot P(\mathcal{T}_{y_c} | \mathcal{T}_v, \mathcal{T}_l). \tag{2}$$

Based on the factorized formulation in Eq. (2), we design a curriculum-guided supervised fine-tuning strategy that progressively guides the multimodal large model to learn the corresponding mappings from intermediate reasoning processes, including scene understanding, key object detection, and root-cause analysis, to the final decision.

Assumption 3.1. *The CoT reasoning process in long-tail scenarios is correlated with the final decision outcome.*

Based on Assumption 3.1 and the conditional probability factorization in Eq. (2), we derive the optimization objective for the cross-entropy loss of curriculum-guided supervised fine-tuning, which is formally stated in Corollary 3.1.

Corollary 3.1. *(CG-SFT) Given that Assumption 3.1 holds, the optimization objective of the curriculum-guided supervised fine-tuning strategy can be formulated from the conditional probability factorization in Eq.* (2) *as follows:*

$$\mathcal{L}_s(\theta) = -\mathbb{E}_{(\mathcal{T}_v, \mathcal{T}_l, \mathcal{T}_{y_d}, \mathcal{T}_{y_c}) \sim \mathcal{D}} \left[\frac{1}{|\mathcal{T}_{y_d}| + |\mathcal{T}_{y_c}|} \left(\sum_{m=1}^{|\mathcal{T}_{y_c}|} \log P_\theta(\mathcal{T}_{y_c,m} \mid \mathcal{T}_v, \mathcal{T}_l, \mathcal{T}_{y_c,<m}) + \sum_{k=1}^{|\mathcal{T}_{y_d}|} \log P_\theta(\mathcal{T}_{y_d,k} \mid \mathcal{T}_{y_c}, \mathcal{T}_v, \mathcal{T}_l, \mathcal{T}_{y_d,<k}) \right) \right]. \tag{3}$$

Based on Corollary 3.1, we adopt a curriculum-guided supervised fine-tuning paradigm in the SFT training process. Concretely, we first pre-train the model using scene understanding and reasoning data, and on this basis, we continue to learn complex reasoning and decision with the full dataset.

3.2 Reasoning-Supervised Policy Optimization algorithm

Our CG-SFT quickly adapts the model to the output distributions of CoT reasoning and decisions via cross-entropy loss. To further boost reasoning robustness and decision accuracy in long-tail scenarios, we introduce a RL strategy guided by reward signals. As shown in Eq. (2) and Assumption 3.1, the final decision is strongly correlated with the CoT reasoning process, requiring both the CoT trajectory and final decision to be considered for policy updates during RL fine-tuning. Inspired by GRPO and GSPO, we optimize decision outputs via reward maximization and compute policy gradients using group-wise relative advantages. Additionally, we optimize CoT reasoning via cross-entropy loss to select reliable sequences highly aligned with decision outcomes. Based on these insights, we propose a Reasoning-Supervised Policy Optimization (RSPO) algorithm, formulated as follows:

$$\mathcal{L}(\theta) = \mathcal{L}_r(\theta) + \lambda \mathcal{L}_c(\theta), \tag{4}$$

where $\mathcal{L}$ denotes the overall loss of our RSPO algorithm; $\mathcal{L}_r$ represents the reward-maximization loss for decision outputs, with the reward defined via rule-based constraints on structured data; $\mathcal{L}_c$ is the cross-entropy loss for the CoT reasoning process; and λ is the weighting coefficient for $\mathcal{L}_c$.

The reward-maximization loss for the decision output is designed to guide structured decision information through rule-based rewards. Since VLMs involve high-dimensional observations and actions, introducing additional critic networks, would increase training costs, and estimation errors from these networks may further affect the accuracy of the advantage function. Therefore, we adopt the relative rather than absolute advantages of sampled actions. Moreover, as this component focuses only on the decision, we measure the distributional discrepancy between the current and old policies using sequence-level importance sampling probabilities that match the reward granularity, rather than token-level importance sampling probabilities. Based on the above analysis, the reward-maximization loss for the decision output can be formulated as follows:

$$\mathcal{L}_r(\theta) = \mathbb{E}_{\mathcal{T}_x \sim D, \{\mathcal{T}_{y_d^i}\}_{i=1}^{G} \sim \pi_{\theta_{\text{old}}}(\cdot|\mathcal{T}_x)} \left[\frac{1}{G} \sum_{i=1}^{G} \min \left(s_i(\theta)\hat{A}_i, \text{clip}(s_i(\theta), 1-\epsilon, 1+\epsilon)\hat{A}_i \right) - \beta D_{\text{KL}}[\pi_\theta \,\|\, \pi_{\text{ref}}] \right], \tag{5}$$

where G denotes the number of samples within each group, ϵ is the clipping parameter that controls the magnitude of policy updates, and β is the KL-divergence penalty coefficient used to balance performance and training stability. For notational convenience, we define $\mathcal{T}_x = [\mathcal{T}_v, \mathcal{T}_l]$. $\hat{A}_i$ denotes the relative advantage of the i-th sample in the group, and $s_i(\theta)$ denotes the sequence-level importance sampling ratio. The sequence-level importance sampling ratio $s_i(\theta)$ and the relative advantage $\hat{A}_i$ can be computed as follows:

$$\hat{A}_i = \frac{r_i - \frac{1}{G}\sum_{j=1}^{G} r_j}{\sqrt{\frac{1}{G}\sum_{j=1}^{G} \left(r_j - \frac{1}{G}\sum_{k=1}^{G} r_k \right)^2}}, \; s_i(\theta) = \exp\left(\frac{1}{|\mathcal{T}_{y_d^i}|} \sum_{t=1}^{|\mathcal{T}_{y_d^i}|} \log \frac{\pi_\theta(\mathcal{T}_{y_d^i,t} \mid \mathcal{T}_X, \mathcal{T}_{y_d^i,<t})}{\pi_{\theta_{\text{old}}}(\mathcal{T}_{y_d^i,t} \mid \mathcal{T}_X, \mathcal{T}_{y_d^i,<t})} \right). \tag{6}$$

Since constraining the magnitude of policy updates can help stabilize policy training to some extent, while an additional KL-divergence penalty may restrict policy optimization, we remove the KL-divergence term when computing the reward-maximization loss for the decision output ($\beta = 0$). The resulting loss function is therefore defined as follows:

$$\mathcal{L}_r(\theta) = \mathbb{E}_{\mathcal{T}_x \sim D, \{\mathcal{T}_{y_d^i}\}_{i=1}^{G} \sim \pi_{\theta_{\text{old}}}(\cdot|\mathcal{T}_x)} \left[\frac{1}{G} \sum_{i=1}^{G} \min \left(s_i(\theta)\hat{A}_i, \text{clip}(s_i(\theta), 1-\epsilon, 1+\epsilon)\hat{A}_i \right) \right]. \tag{7}$$

The objective of the log-likelihood loss for CoT reasoning is to guide the policy to generate CoT reasoning trajectories that are consistent with the ground-truth decisions. Therefore, the CoT reasoning loss is defined as follows:

$$\mathcal{L}_c(\theta) = \mathbb{E}_{\mathcal{T}_x \sim D, \{\mathcal{T}_{y_c^i}\}_{i=1}^{G} \sim \pi_{\theta_{\text{old}}}(\cdot|\mathcal{T}_x)} \left[\frac{1}{G} \sum_{i=1}^{G} \left(-\frac{1}{T_i} \sum_{t=1}^{T_i} \log \pi_\theta \left(\mathcal{T}_{y_c^i,t} \mid \mathcal{T}_x, \mathcal{T}_{y_c^i,<t} \right) \right) \right], \tag{8}$$

where T_i denotes the token length of the sampled rollout trajectory. Based on the above analysis, the optimization objective of the chain-of-thought supervised optimization strategy can be defined as stated in Corollary 3.2.

Corollary 3.2. *(RSPO) Under Assumption 3.1, the reasoning-supervised optimization algorithm adopts reward-guided optimization for decision outputs and imposes cross-entropy loss on the*

chain-of-thought reasoning process. The corresponding loss function is defined as follows:

$$\mathcal{L}(\theta) = \underbrace{\mathbb{E}_{\mathcal{T}_x \sim D, \{\mathcal{T}_{y_d^i}\}_{i=1}^G \sim \pi_{\theta_{old}}(\cdot|\mathcal{T}_x)} \left[\frac{1}{G} \sum_{i=1}^{G} \min \left(s_i(\theta)\hat{A}_i, \mathit{clip}(s_i(\theta), 1-\epsilon, 1+\epsilon)\hat{A}_i \right) \right]}_{\mathcal{L}_r(\theta)} + \lambda \cdot \underbrace{\mathbb{E}_{\mathcal{T}_x \sim D, \{\mathcal{T}_{y_c^i}\}_{i=1}^G \sim \pi_{\theta_{old}}(\cdot|\mathcal{T}_x)} \left[\frac{1}{G} \sum_{i=1}^{G} \left(-\frac{1}{|\mathcal{T}_{y_c^i}|} \sum_{t=1}^{|\mathcal{T}_{y_c^i}|} \log \pi_\theta \left(\mathcal{T}_{y_c^i,t} \mid \mathcal{T}_x, \mathcal{T}_{y_c^i,<t} \right) \right) \right]}_{\mathcal{L}_c(\theta)}. \tag{9}$$

Corollary 3.2 provides the optimization objectives for maximizing the decision reward and minimizing the cross-entropy of the CoT reasoning process.

4 Related Work

VLA for Autonomous Driving. VLA models unify perception, language understanding, and action generation, achieving notable success in robotics [4, 5, 28, 65]. This paradigm has been progressively adopted in end-to-end autonomous driving, evolving from LLM-based policies [7, 38, 52] to VLM-based approaches with enhanced multi-modal scene understanding [45, 63]. Planning methods have also diversified beyond trajectory regression, including diffusion-based generation [27], cognition distillation [30], and world-model pretraining [29, 34, 46]. To improve reasoning in complex scenarios, CoT [48]has been integrated into driving VLAs: EMMA [26] pairs trajectory generation with interpretable reasoning; AutoVLA [64] adaptively adjusts reasoning depth by scene complexity; FSDrive [57] uses world-model-predicted future states as intermediate reasoning representations. However, these methods remain confined to supervised learning on offline expert data, limiting generalization to long-tail scenarios and lacking closed-loop feedback for policy improvement.

Reinforcement Learning for VLM and VLA. Recent work applies reinforcement learning (RL) post-training to enhance driving reasoning and planning robustness. AutoDrive-R^2 [55] combines GRPO with physics-grounded rewards for self-reflection; AdaThinkDrive [37] uses adaptive rewards for dynamic reasoning depth selection; Reasoning-VLA [58] and Drive-R1 [31] strengthen planning via action-query interaction and multi-dimensional rewards; Alpamayo-R1 [47] optimizes reasoning quality and trajectory safety through GRPO. However, a fundamental challenge persists: *the lack of consistency optimization between reasoning and decisions*. Most methods apply rewards solely to final answers, leaving reasoning chains unsupervised and prone to templated degeneration. Sequence-level approaches (e.g., WAM-Diff [51] with GSPO) impose uniform constraints over entire outputs, increasing entropy and hindering decision optimality. Auxiliary reward models [60, 61] can jointly assess reasoning and decisions but are bounded by their own accuracy and incur prohibitive overhead.

We propose a RSPO algorithm that jointly maximizes decision rewards and minimizes reasoning cross-entropy, enforcing reasoning-decision consistency without additional reward models, improving both robustness and accuracy on long-tail scenarios.

5 Experiment

We evaluate the proposed strategy on multiple long-tail datasets under shared settings, analyze individual components and hyperparameters, and construct verifiable reasoning-and-decision datasets from open-source benchmarks because no ready-to-use benchmark exists for this setting. The evaluation covers structured decision metrics, CoT-related reasoning behavior, reinforcement-learning dynamics, and downstream trajectory-prediction errors. This design tests not only the accuracy of the final decision, but also whether reasoning supervision improves the robustness of the policy under long-tail conditions.

5.1 Dataset Construction

We construct long-tail autonomous driving reasoning and decision-making datasets from three sources: *CODA-LM* [6] for real-world long-tail road scenes with multi-mode decisions and CoT reasoning, *Impromptu* [9] for unstructured scenarios involving vulnerable road users and traffic lights, and *RoadWork* [12] for construction-zone decisions. These datasets cover complementary long-tail factors, including hazardous-source localization, vulnerable-road-user reasoning, traffic-signal

Table 1: Qualitative cases and quantitative performance on the CODA-LM autonomous driving reasoning and decision task. The left panel shows representative high-scoring long-tail cases with input images, localized risk sources, and RSPO outputs. In the table, *Event* indicates whether a scenario is a long-tail safety-critical case; *Category* is the classification metric for hazardous long-tail scenarios; *Bbox* measures hazardous-source localization; *Long* and *Lat* evaluate longitudinal and lateral driving suggestions, respectively; and *Mean* reports the average over all metrics.



Method	Event	Category	Bbox	Long	Lat	Mean
Zero Shot						
GPT-5.4 [42]	61.5	32.4	22.8	69.7	57.6	55.3
Gemini-3-Flash [13]	63.1	39.3	37.0	65.4	55.2	55.8
Qwen3.6-Plus [39]	68.0	33.8	42.6	44.0	51.1	49.2
Doubao-Seed [40]	59.7	38.9	37.5	59.5	54.4	53.1
MiMo-V2-Omni [50]	58.7	34.0	41.0	55.2	45.4	48.3
Gemma-4-31B-It [14]	65.6	35.9	32.3	64.8	55.4	55.4
Qwen3.5-397B-A17B [44]	61.9	34.0	40.4	54.4	53.6	51.0
Qwen3.5-27B [44]	58.3	32.8	38.4	53.4	52.8	49.3
GRPO						
InternVL-8B [8]	81.7	66.2	61.3	59.9	53.0	65.2
MiMo-VL-7B [56]	80.0	61.9	57.7	60.6	53.1	63.9
Qwen3-VL-4B [2]	78.0	60.3	59.4	60.4	55.9	63.7
Qwen2.5-VL-3B [3]	77.8	60.1	54.9	63.0	53.0	63.5
Qwen2.5-VL-7B [3]	77.0	58.5	56.7	59.7	54.7	62.5
Ours (Qwen3-VL-4B)						
CG-SFT	*79.2*	*65.4*	*62.8*	*60.2*	*57.9*	*65.1*
CG-SFT+RSPO	**85.7**	**71.1**	**64.8**	**64.5**	**61.6**	**69.5**

ambiguity, and construction-zone constraints. All datasets combine large-model pre-annotation, human verification, and re-annotation to produce verifiable structured labels and reasoning traces. Detailed pipelines, input/output definitions, and evaluation protocols are provided in Appendix A.1.1.

5.2 Experimental Settings and Baselines

Our method uses Qwen3-VL-4B as the base model and compares against scale-comparable VLMs, including InternVL-8B, MiMO-VL-7B, Qwen2.5-VL-3B, and Qwen2.5-VL-7B. We evaluate standard SFT and three RL post-training baselines, GRPO, GSPO, and DAPO, under the same training configuration to isolate the effect of the optimization objective. For dataset characterization, we also include closed- and open-source models prompted with a small number of expert demonstrations; detailed settings and reward definitions are provided in Appendix A.1.2 and A.1.3.

5.3 Performance Comparison Experiment

We evaluate curriculum-guided SFT and reasoning-supervised reinforcement learning on CODA-LM, Impromptu, and RoadWork. The comparison includes zero-shot prompted models, SFT baselines, GRPO-enhanced models, and our CG-SFT+RSPO strategy. Across datasets, the metrics measure both perception-related fields and structured longitudinal/lateral decisions, so the reported mean scores reflect reasoning-to-action consistency rather than a single classification target.

Performance Comparison on CODA-LM. As shown in Table 1, our CG-SFT+RSPO strategy is evaluated on the CODA-LM autonomous driving reasoning and decision-making task against zero-shot and GRPO-based baselines. While zero-shot models remain limited on fine-grained category and bounding-box metrics, GRPO-enhanced VLMs improve the mean score to 62.5–65.2. Our method achieves the best mean score of 69.5 and the top performance across all metrics. The gains are especially clear on Event, Category, Bbox, and Lateral, showing that RSPO improves risk recognition, hazard localization, and structured driving suggestions in long-tail safety-critical scenes. This pattern suggests that supervising the reasoning trajectory complements reward-based decision optimization, rather than merely increasing final-answer accuracy.

Performance Comparison on Impromptu and RoadWork. As shown in Table 2, zero-shot models achieve mean scores of 65.0–70.0 on Impromptu and 45.2–52.5 on RoadWork, while GRPO baselines reach 94.2–96.1 and 80.2–83.3, respectively. CG-SFT is already strong on Impromptu (96.7), and RSPO further obtains the best means on both tasks: 96.9 on Impromptu and 84.6 on RoadWork. It

Table 2: Qualitative cases and quantitative performance on the Impromptu and RoadWork autonomous driving reasoning and decision-making tasks. The left panel shows representative Impromptu and RoadWork cases with input images and model outputs. For Impromptu, VRU and Light denote vulnerable road user recognition and traffic light recognition, respectively; Longitudinal and Lateral evaluate longitudinal and lateral driving suggestions; Mean reports the average over all metrics.



Method	Impromptu					RoadWork		
	VRU	Light	Long	Lat	Mean	Long	Lat	Mean
Zero Shot								
GPT-5.4 [42]	88.4	90.6	41.4	46.6	66.8	29.6	75.4	52.5
Gemini-3-Flash [13]	91.0	87.0	45.0	44.8	67.0	27.8	78.6	53.2
Qwen3.6-Plus [39]	93.4	95.4	42.2	34.0	66.3	17.0	79.0	48.0
Doubao-Seed [40]	88.0	95.2	42.3	41.9	66.8	12.4	79.8	46.1
MiMo-V2-Omni [50]	92.2	94.8	45.2	47.6	70.0	24.2	68.8	46.5
Gemma-4-31B-It [14]	91.1	93.9	44.2	46.2	68.9	15.1	75.2	45.2
Qwen3.5-397B-A17B [44]	91.7	94.2	43.9	37.7	66.9	17.0	78.0	47.5
Qwen3.5-27B [44]	87.1	88.8	40.1	44.1	65.0	37.9	57.4	47.7
GRPO								
InternVL-8B [8]	92.4	95.5	94.7	94.5	94.2	77.6	85.4	81.5
MiMo-VL-7B [56]	94.9	96.1	81.5	81.5	88.5	80.1	86.5	83.3
Qwen3-VL-4B [2]	95.7	97.3	96.6	95.9	96.4	81.1	87.5	84.3
Qwen2.5-VL-3B [3]	93.8	97.0	95.4	94.9	95.3	76.4	84.1	80.2
Qwen2.5-VL-7B [3]	95.1	96.0	96.7	96.7	96.1	80.1	86.7	83.4
Ours (Qwen3-VL-4B)								
CG-SFT	*96.4*	*97.0*	*96.7*	*96.6*	*96.7*	–	–	–
CG-SFT+RSPO	**96.9**	**97.3**	**96.9**	**96.4**	**96.9**	**81.3**	**87.9**	**84.6**

improves over the strongest GRPO baseline by 0.48 and 0.36 points on Impromptu and RoadWork, respectively. Improvements are smaller than on CODA-LM since GRPO baselines are close to saturation for multiple decision domains, though RSPO still consistently boosts the overall mean.

5.4 Comparison of Different RL Algorithms

Figure 4 compares RL algorithms under the same CG-SFT initialization. DAPO and GSPO improve faster early in training, whereas RSPO shows steadier growth and converges to a markedly higher final reward. This behavior suggests that preserving CoT supervision during policy optimization slows early reward chasing but yields better final alignment between reasoning and decisions, which is consistent with the benchmark gains above.

5.5 Comparative Evaluation of Downstream Trajectory Planning

For downstream planning, we curate a long-tail AR1 trajectory-prediction subset in which each sample uses four front-view frames and ego-motion context to predict future waypoints; screening, splits, and metrics are detailed in Appendix A.1.1. Table 3 shows that adding decision output reduces SFT ADE@8s/FDE@8s from 3.429/10.111 to 3.246/9.512, and also improves the GRPO setting from 3.261/9.578 to 3.168/9.183. RSPO further achieves the lowest ADE@8s/FDE@8s errors (2.993/8.585). These results indicate that reasoning-decision supervision transfers beyond discrete decisions and improves trajectory prediction under long-tail conditions.

5.6 Ablation Experiment

To evaluate the contributions of the proposed curriculum-guided SFT strategy and the chain-of-thought supervised reinforcement learning strategy, we conduct ablation studies on each component. In addition, we perform hyperparameter ablations to identify the optimal configuration.

Ablation on Curriculum-Guided SFT and RSPO. As shown in Figure 5a, we conduct ablation studies on the CODA-LM task to evaluate our Curriculum-Guided SFT (CG-SFT) and RSPO against the standard SFT baseline. CG-SFT (blue bars) delivers an overall mean improvement of +1.82% over SFT, with notable gains in Category (+4.48%) and Lateral (+3.46%), despite a minor drop in Longitudinal (-4.48%). RSPO (orange bars) achieves a more substantial mean improvement of +6.25%, outperforming SFT on all metrics and yielding the largest gain in Category (+10.18%). These results validate that both curriculum-guided fine-tuning and reasoning-supervised reinforcement

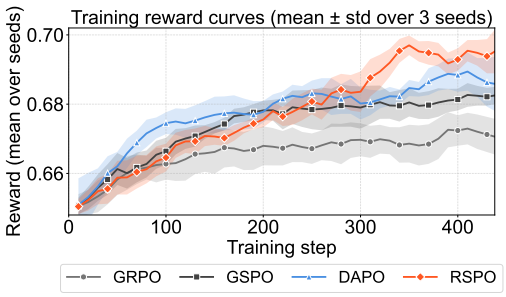


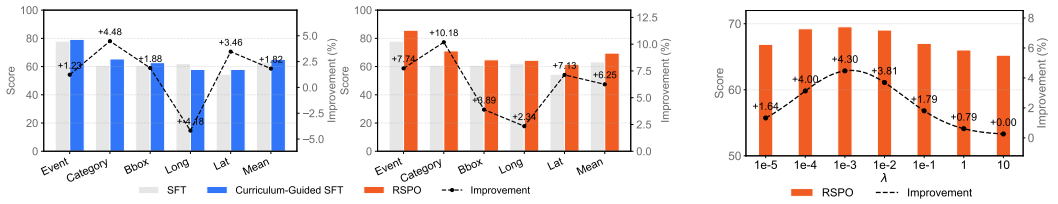
Figure 4: Reward curves on the test set during training for different RL algorithms built upon CG-SFT evaluated on the CODA-LM dataset.

Table 3: Experimental results on the AR1 task: L2 errors and ADE, FDE under different strategies.

Method	L2@1S	L2@2S	L2@3S	ADE@8s	FDE@8s
SFT					
Trajs	0.081	0.267	0.588	3.429	10.111
Trajs+Dec	**0.080**	**0.260**	0.568	3.246	9.512
SFT+GRPO					
Trajs	0.081	0.263	0.573	3.261	9.578
Trajs+Dec	0.085	0.268	0.576	3.168	9.183
SFT+RSPO(Ours)					
Trajs+Dec	0.083	0.264	**0.563**	**2.993**	**8.585**

learning (RSPO) effectively enhance long-tail reasoning performance, with RSPO providing more consistent and robust improvements across all evaluated metrics.

Ablation on the Hyperparameter λ. As shown in Figure 5b, we conduct an ablation study on the hyperparameter λ of our RSPO algorithm for the CODA-LM task. The results demonstrate that the performance of RSPO's first increases and then decreases as λ increases from 10^{-5} to 10. Optimal performance (mean score ≈ 69, improvement $+4.30\%$) is achieved at $\lambda = 10^{-3}$, effectively balancing the reward signal and the stability of the policy. As λ exceeds 10^{-3}, performance and improvement gradually decrease, and gain disappear completely at $\lambda = 10$. This confirms that a properly tuned λ is critical to leveraging reasoning supervision for performance gains, while an overly large λ impairs model stability and generalization.



(a) Comparison of CG-SFT and RSPO against the standard SFT baseline. (b) Hyperparameter λ for RSPO.

Figure 5: Ablation studies. (a) CG-SFT denotes curriculum-guided supervised fine-tuning, while RSPO denotes reinforcement fine-tuning initialized from the standard SFT model. (b) Mean evaluation metrics on the CODA-LM task under different λ values.

6 Conclusion

In this paper, we studied reasoning and decision-making for long-tail autonomous driving scenarios and proposed RSPO, a reasoning-supervised policy optimization framework. RSPO first uses curriculum-guided supervised fine-tuning to learn structured CoT reasoning and driving decisions, and then jointly optimizes decision rewards and CoT supervision during reinforcement fine-tuning. Experiments on CODA-LM, Impromptu, and RoadWork show that RSPO achieves the best mean scores of 69.53, 96.86, and 84.63, respectively, outperforming strong GRPO baselines by 4.36, 0.48, and 0.36 points. Moreover, downstream experiments on AR1 trajectory planning further validate its effectiveness, where RSPO obtains the lowest ADE@8s/FDE@8s errors of 2.993/8.585.

Limitations. Although RSPO achieves promising results, our experiments are mainly conducted on offline datasets derived from open-source benchmarks, and its effectiveness in closed-loop simulation or real-world deployment remains to be further verified. In addition, the current reward design relies on structured annotations and rule-based evaluation, which may limit scalability to more diverse driving behaviors and highly ambiguous scenarios. Future work will extend RSPO to closed-loop evaluation and more fine-grained safety-aware reward modeling.

Acknowledgments and Disclosure of Funding

This work is supported by the Beijing Nova Program under Grant No. 202604841209. We are grateful to the authors of publicly available datasets for making their valuable data accessible, which has greatly benefited our research. We also thank the anonymous reviewers for their perceptive and constructive feedback, which has significantly improved the quality of this manuscript.

References

[1] Arash Ahmadian, Chris Cremer, Matthias Gallé, Marzieh Fadaee, Julia Kreutzer, Olivier Pietquin, Ahmet Üstün, and Sara Hooker. Back to basics: Revisiting reinforce-style optimization for learning from human feedback in llms. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12248–12267, 2024.

[2] Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, et al. Qwen3-vl technical report, 2025. URL `https://arxiv.org/abs/2511.21631`.

[3] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, et al. Qwen2.5-vl technical report, 2025. URL `https://arxiv.org/abs/2502.13923`.

[4] Johan Bjorck, Fernando Castañeda, Nikita Cherniadev, Xingye Da, Runyu Ding, Linxi Fan, Yu Fang, Dieter Fox, Fengyuan Hu, Spencer Huang, et al. Gr00t n1: An open foundation model for generalist humanoid robots. *arXiv preprint arXiv:2503.14734*, 2025.

[5] Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, et al. A vision-language-action flow model for general robot control. *arXiv preprint arXiv:2410.24164*, 2024.

[6] Kai Chen, Yanze Li, Wenhua Zhang, Yanxin Liu, Pengxiang Li, Ruiyuan Gao, Lanqing Hong, Meng Tian, Xinhai Zhao, Zhenguo Li, et al. Automated evaluation of large vision-language models on self-driving corner cases. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 7817–7826. IEEE, 2025.

[7] Long Chen, Oleg Sinavski, Jan Hünermann, Alice Karnsund, Andrew James Willmott, Danny Birch, Daniel Maund, and Jamie Shotton. Driving with llms: Fusing object-level vector modality for explainable autonomous driving. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 14093–14100. IEEE, 2024.

[8] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 24185–24198, 2024.

[9] Haohan Chi, Huan-ang Gao, Ziming Liu, Jianing Liu, Chenyu Liu, Jinwei Li, Kaisen Yang, Yangcheng Yu, Zeda Wang, Wenyi Li, et al. Impromptu vla: Open weights and open data for driving vision-language-action models. *arXiv preprint arXiv:2505.23757*, 2025.

[10] Bowen Feng, Zhiting Mei, Julian Ost, Filippo Ghilotti, Baiang Li, Roger Girgis, Anirudha Majumdar, and Felix Heide. Verdi: Vlm-embedded reasoning for autonomous driving. *arXiv preprint arXiv:2505.15925*, 2025.

[11] Chang Gao, Chujie Zheng, Xiong-Hui Chen, Kai Dang, Shixuan Liu, Bowen Yu, An Yang, Shuai Bai, Jingren Zhou, and Junyang Lin. Soft adaptive policy optimization. *arXiv preprint arXiv:2511.20347*, 2025.

[12] Anurag Ghosh, Shen Zheng, Robert Tamburo, Khiem Vuong, Juan Alvarez-Padilla, Hailiang Zhu, Michael Cardei, Nicholas Dunn, Christoph Mertz, and Srinivasa G Narasimhan. Roadwork: A dataset and benchmark for learning to recognize, observe, analyze and drive through work zones. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6132–6142, 2025.

[13] Google DeepMind. Gemini. https://deepmind.google/models/gemini/, 2026.

[14] Google DeepMind. Gemma 4. https://deepmind.google/models/gemma/gemma-4/, 2026.

[15] Jiayi Guan, Guang Chen, Jin Huang, Zhijun Li, Lu Xiong, Jing Hou, and Alois Knoll. A discrete soft actor-critic decision-making strategy with sample filter for freeway autonomous driving. *IEEE Transactions on Vehicular Technology*, 72(2):2593–2598, 2022.

[16] Jiayi Guan, Li Shen, Ao Zhou, Lusong Li, Han Hu, Xiaodong He, Guang Chen, and Changjun Jiang. Poce: Primal policy optimization with conservative estimation for multi-constraint offline reinforcement learning. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 26233–26243. IEEE, 2024.

[17] Jiayi Guan, Tianle Zhang, Li Shen, Ruiqi Zhang, Ao Zhou, Lusong Li, Guai Chen, Mengjie Li, Alois Knoll, Xiaodong He, et al. Cdcp: Conditional diffusion model with contextual prompts for multi-task offline safe reinforcement learning. *arXiv preprint arXiv:2607.03903*, 2026.

[18] Yufei Guo, Yinan Wu, Haoran Duan, Guiguang Ding, and Jungong Han. Phaselora: Control-regime-conditioned low-rank adaptation for continuous-action vision-language-action policies. *arXiv preprint arXiv:2608.15285*, 2026.

[19] Xiaoshuai Hao, Lei Zhou, Zhijian Huang, Zhiwen Hou, Yingbo Tang, Lingfeng Zhang, Guang Li, Zheng Lu, Shuhuai Ren, Xianhui Meng, et al. Mimo-embodied: X-embodied foundation model technical report. *arXiv preprint arXiv:2511.16518*, 2025.

[20] Minghui Hou, Wei-Hsing Huang, Shaofeng Liang, Daizong Liu, Tai-Hao Wen, Gang Wang, Runwei Guan, and Weiping Ding. Mmdrive: Interactive scene understanding beyond vision with multi-representational fusion. *arXiv preprint arXiv:2512.13177*, 2025.

[21] Jian Hu, Jason Klein Liu, Haotian Xu, and Wei Shen. Reinforce++: Stabilizing critic-free policy optimization with global normalization. *arXiv preprint arXiv:2501.03262*, 2025.

[22] Xin Hu, Taotao Jing, Renran Tian, and Zhengming Ding. Vlms guided interpretable decision making in autonomous driving. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 4494–4503, 2026.

[23] Wenhui Huang, Songyan Zhang, Qihang Huang, Zhidong Wang, Zhiqi Mao, Collister Chua, Zhan Chen, Long Chen, and Chen Lv. Automot: A unified vision-language-action model with asynchronous mixture-of-transformers for end-to-end autonomous driving. *arXiv preprint arXiv:2603.14851*, 2026.

[24] Zhijian Huang, Tao Tang, Shaoxiang Chen, Sihao Lin, Zequn Jie, Lin Ma, Guangrun Wang, and Xiaodan Liang. Making large language models better planners with reasoning-decision alignment. In *European Conference on Computer Vision*, pages 73–90. Springer, 2024.

[25] Zhijian Huang, Chengjian Feng, Feng Yan, Baihui Xiao, Zequn Jie, Yujie Zhong, Xiaodan Liang, and Lin Ma. Robotron-drive: All-in-one large multimodal model for autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8011–8021, 2025.

[26] 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.

[27] Anqing Jiang, Yu Gao, Zhigang Sun, Yiru Wang, Jijun Wang, Jinghao Chai, Qian Cao, Yuweng Heng, Hao Jiang, Yunda Dong, et al. Diffvla: Vision-language guided diffusion planning for autonomous driving. *arXiv preprint arXiv:2505.19381*, 2025.

[28] Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, et al. Openvla: An open-source vision-language-action model. *arXiv preprint arXiv:2406.09246*, 2024.

[29] Yingyan Li, Shuyao Shang, Weisong Liu, Bing Zhan, Haochen Wang, Yuqi Wang, Yuntao Chen, Xiaoman Wang, Yasong An, Chufeng Tang, et al. Drivevla-w0: World models amplify data scaling law in autonomous driving. *arXiv preprint arXiv:2510.12796*, 2025.

[30] Yongkang Li, Kaixin Xiong, Xiangyu Guo, Fang Li, Sixu Yan, Gangwei Xu, Lijun Zhou, Long Chen, Haiyang Sun, Bing Wang, et al. Recogdrive: A reinforced cognitive framework for end-to-end autonomous driving. *arXiv preprint arXiv:2506.08052*, 2025.

[31] Yue Li, Meng Tian, Dechang Zhu, Jiangtong Zhu, Zhenyu Lin, Zhiwei Xiong, and Xinhai Zhao. Drive-r1: Bridging reasoning and planning in vlms for autonomous driving with reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 6708–6716, 2026.

[32] Zhuofan Li, Hongkun Yang, Zhenyang Chen, Yangxuan Chen, Chaojian Li, et al. From inference efficiency to embodied efficiency: Revisiting efficiency metrics for vision-language-action models. *arXiv preprint arXiv:2603.19131*, 2026.

[33] Haichao Liu, Haoren Guo, Pei Liu, Benshan Ma, Yuxiang Zhang, Jun Ma, and Tong Heng Lee. Vlm-udmc: Vlm-enhanced unified decision-making and motion control for urban autonomous driving. *arXiv preprint arXiv:2507.15266*, 2025.

[34] Lin Liu, Ziying Song, Caiyan Jia, Hangjun Ye, Xiaoshuai Hao, Long Chen, et al. Driveworld-vla: Unified latent-space world modeling with vision-language-action for autonomous driving. *arXiv preprint arXiv:2602.06521*, 2026.

[35] Jinghui Lu, Jiayi Guan, Zhijian Huang, Jinlong Li, Guang Li, Lingdong Kong, Yingyan Li, Han Wang, Shaoqing Xu, Yuechen Luo, et al. Onevl: One-step latent reasoning and planning with vision-language explanation. *arXiv e-prints*, pages arXiv–2604, 2026.

[36] Yuhang Lu, Jiadong Tu, Yuexin Ma, and Xinge Zhu. Real-ad: Towards human-like reasoning in end-to-end autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 27783–27793, 2025.

[37] Yuechen Luo, Fang Li, Shaoqing Xu, Zhiyi Lai, Lei Yang, Qimao Chen, Ziang Luo, Zixun Xie, Shengyin Jiang, Jiaxin Liu, et al. Adathinkdrive: Adaptive thinking via reinforcement learning for autonomous driving. *arXiv preprint arXiv:2509.13769*, 2025.

[38] Jiageng Mao, Yuxi Qian, Junjie Ye, Hang Zhao, and Yue Wang. Gpt-driver: Learning to drive with gpt. In *NeurIPS 2023 Foundation Models for Decision Making Workshop*, 2024.

[39] Qwen Team. Qwen3.6-Plus: Towards real world agents, April 2026. URL `https://qwen.ai/blog?id=qwen3.6`.

[40] Bytedance Seed. Seed2. 0 model card: Towards intelligence frontier for real-world complexity. Technical report, Technical report, Bytedance, 2025. URL https://lf3-static. bytednsdoc. com . . . , 2026.

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

[42] Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, et al. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267*, 2025.

[43] Shounak Sural, Ragunathan Raj Rajkumar, et al. Contextvlm: Zero-shot and few-shot context understanding for autonomous driving using vision language models. In *2024 IEEE 27th International Conference on Intelligent Transportation Systems (ITSC)*, pages 468–475. IEEE, 2024.

[44] Qwen Team. Qwen3.5: Accelerating productivity with native multimodal agents, February 2026. URL `https://qwen.ai/blog?id=qwen3.5`.

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

[46] Guoqing Wang, Pin Tang, Xiangxuan Ren, Guodongfang Zhao, Bailan Feng, and Chao Ma. Learning vision-language-action world models for autonomous driving. *arXiv preprint arXiv:2604.09059*, 2026.

[47] Yan Wang, Wenjie Luo, Junjie Bai, Yulong Cao, Tong Che, Ke Chen, Yuxiao Chen, Jenna Diamond, Yifan Ding, Wenhao Ding, et al. Alpamayo-r1: Bridging reasoning and action prediction for generalizable autonomous driving in the long tail. *arXiv preprint arXiv:2511.00088*, 2025.

[48] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

[49] Jing Wu, Jianhua Wu, Jiayi Guan, Jiahong Chen, Jinghui Lu, Hangjun Ye, Bingzhao Gao, and Long Chen. Spatiolm: Towards general physical spatial intelligence in vision-language models. *arXiv preprint arXiv:2608.01899*, 2026.

[50] Xiaomi MiMo Team. Mimo-v2-omni. `https://mimo.xiaomi.com/mimo-v2-omni`, 2026.

[51] Mingwang Xu, Jiahao Cui, Feipeng Cai, Hanlin Shang, Zhihao Zhu, Shan Luan, Yifang Xu, Neng Zhang, Yaoyi Li, Jia Cai, et al. Wam-diff: A masked diffusion vla framework with moe and online reinforcement learning for autonomous driving. *arXiv preprint arXiv:2512.11872*, 2025.

[52] 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*, 9(10):8186–8193, 2024.

[53] Ruoyu Yao, Yubin Wang, Haichao Liu, Rui Yang, Zengqi Peng, Lei Zhu, and Jun Ma. Calmm-drive: Confidence-aware autonomous driving with large multimodal model. *arXiv preprint arXiv:2412.04209*, 2024.

[54] Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, et al. Dapo: An open-source llm reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*, 2025.

[55] Zhenlong Yuan, Chengxuan Qian, Jing Tang, Rui Chen, Zijian Song, Lei Sun, Xiangxiang Chu, Yujun Cai, Dapeng Zhang, and Shuo Li. Autodrive-r^2: Incentivizing reasoning and self-reflection capacity for vla model in autonomous driving. *arXiv preprint arXiv:2509.01944*, 2025.

[56] Zihao Yue, Zhenru Lin, Yifan Song, Weikun Wang, Shuhuai Ren, Shuhao Gu, Shicheng Li, et al. Mimo-vl technical report, 2025. URL `https://arxiv.org/abs/2506.03569`.

[57] Shuang Zeng, Xinyuan Chang, Mengwei Xie, Xinran Liu, Yifan Bai, Zheng Pan, Mu Xu, Xing Wei, and Ning Guo. Futuresightdrive: Thinking visually with spatio-temporal cot for autonomous driving. *arXiv preprint arXiv:2505.17685*, 2025.

[58] Dapeng Zhang, Zhenlong Yuan, Zhangquan Chen, Chih-Ting Liao, Yinda Chen, Fei Shen, Qingguo Zhou, and Tat-Seng Chua. Reasoning-vla: A fast and general vision-language-action reasoning model for autonomous driving. *arXiv preprint arXiv:2511.19912*, 2025.

[59] Jianke Zhang, Yanjiang Guo, Yucheng Hu, Xiaoyu Chen, Xiang Zhu, and Jianyu Chen. Up-vla: A unified understanding and prediction model for embodied agent. *arXiv preprint arXiv:2501.18867*, 2025.

[60] Lunjun Zhang, Arian Hosseini, Hritik Bansal, Mehran Kazemi, Aviral Kumar, and Rishabh Agarwal. Generative verifiers: Reward modeling as next-token prediction. *arXiv preprint arXiv:2408.15240*, 2024.

[61] YiFan Zhang, Xingyu Lu, Xiao Hu, Chaoyou Fu, Bin Wen, Tianke Zhang, Changyi Liu, Kaiyu Jiang, Kaibing Chen, Kaiyu Tang, Haojie Ding, Jiankang Chen, Fan Yang, Zhang Zhang, Tingting Gao, Di ZHANG, Guorui Zhou, and Liang Wang. R1-reward: Training multimodal reward model through stable reinforcement learning. In *The Fourteenth International Conference on Learning Representations*, 2026. URL `https://openreview.net/forum?id=4Ewgw9M2xE`.

[62] Chujie Zheng, Shixuan Liu, Mingze Li, Xiong-Hui Chen, Bowen Yu, Chang Gao, Kai Dang, Yuqiong Liu, Rui Men, An Yang, et al. Group sequence policy optimization. *arXiv preprint arXiv:2507.18071*, 2025.

[63] Xingcheng Zhou, Xuyuan Han, Feng Yang, Yunpu Ma, Volker Tresp, and Alois Knoll. Opendrivevla: Towards end-to-end autonomous driving with large vision language action model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 13782–13790, 2026.

[64] 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. *arXiv preprint arXiv:2506.13757*, 2025.

[65] Brianna Zitkovich, Tianhe Yu, Sichun Xu, Peng Xu, Ted Xiao, Fei Xia, Jialin Wu, Paul Wohlhart, Stefan Welker, Ayzaan Wahid, et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. In *Conference on Robot Learning*, pages 2165–2183. PMLR, 2023.

A Technical Appendices and Supplementary Material

A.1 Additional Experiments

A.1.1 Dataset Details

We construct long-tail autonomous driving reasoning and decision-making datasets from multiple data sources, including CODA-LM, Impromptu, RoadWork, and AR1. Each dataset block below pairs the construction details with its corresponding overview figure, so that the input/output definition, annotation pipeline, and evaluation protocol can be read together.

CODA-LM dataset. CODA-LM is derived from CODA, a real-world dataset targeting long-tail road scenarios, and provides annotations for general perception, regional perception, and driving suggestions. Based on the CODA-LM dataset, we adopt a pipeline of large model pre-annotation, manual verification, and large model re-annotation. On the long-tail scenarios covered by CODA-LM, we newly annotate multiple labels, including long-tail hazardous scenario identification, long-tail scenario category, hazard sources, as well as longitudinal and lateral driving suggestions. Meanwhile, we supplement scene descriptions and Chain-of-Thought (CoT) analysis for the root causes of driving decisions. Since CODA-LM does not provide navigation and ego-vehicle information, multiple plausible driving decisions may exist in the decision-making process. Accordingly, we present multi-mode decision results in our final annotations.

The construction pipeline of our dataset is detailed as follows: (1) We leverage closed-source large language models to automatically annotate scene analysis and initial decisions via prompt engineering; (2) We manually verify the finalized decisions and all structured annotation labels; (3) Guided by the manually validated structured labels, we employ closed-source large language models again to generate scene analysis and reasoning content that conforms to label specifications. Our dataset takes a single-frame front-view image as input, and outputs the following contents: identification of long-tail hazardous scenarios, category of long-tail hazardous scenarios, hazard sources, dual-mode longitudinal and lateral driving decisions, as well as scene descriptions and CoT-based reasoning analysis. In CODA-LM, the scene categories are defined as A: extreme weather or lighting, B: road construction or temporary rerouting, C: abnormal traffic-participant behavior, D: irregular objects or debris, E: complex combined risk, and F: regular safe scene. Built upon CODA-LM, our dataset contains 4702 samples in the training set and 491 samples in the test set. Figure 6 summarizes the CODA-LM data construction flow, input/output fields, and evaluation metrics.

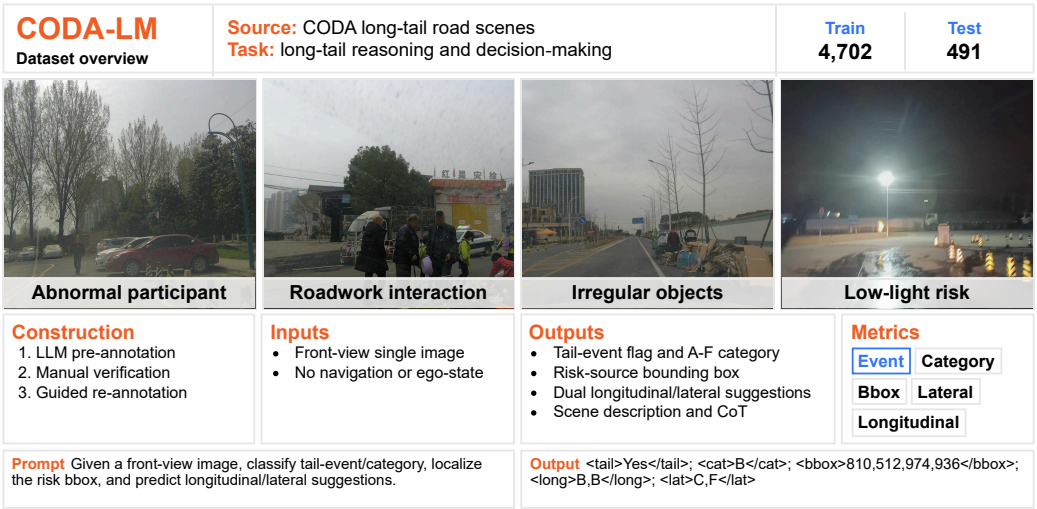


Figure 6: Dataset overview of CODA-LM, including representative long-tail scenes, construction process, input/output fields, and evaluation metrics.

Impromptu dataset. The Impromptu dataset focuses on four typical categories of unstructured long-tail driving scenarios. It selects 80,000 high-quality video clips from eight public datasets, which

are equipped with planning-oriented question-answer annotations and trajectory labels. Based on the original annotations of Impromptu, we construct a high-quality dataset for root-cause reasoning and decision-making, containing 14,004 training samples and 3,808 test samples. This dataset takes single-view imagery, navigation information, and ego-vehicle information as inputs, and outputs the status of vulnerable road users and traffic lights, driving decision results, along with the analysis of scene understanding and decision root causes. Figure 7 provides the corresponding overview of the Impromptu construction process and evaluation protocol.

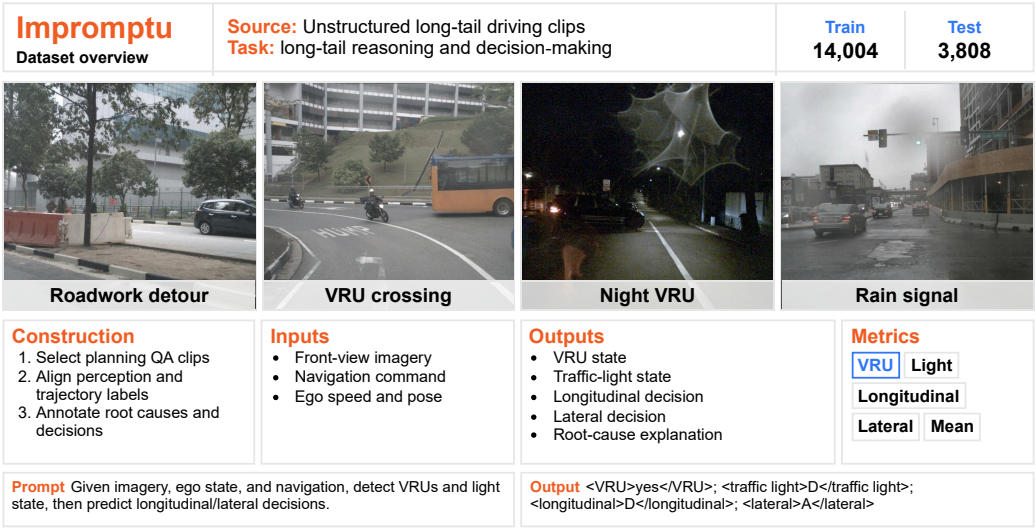


Figure 7: Dataset overview of Impromptu, including representative unstructured scenes, construction process, input/output fields, and evaluation metrics.

RoadWork dataset. RoadWork is a large-scale dataset and benchmark for autonomous driving in road construction zones, covering over 5,000 real-world construction scenarios. Leveraging the original annotations of this dataset, we extract driving decisions and Chain-of-Thought (CoT) analysis for corresponding root causes, and further construct a long-tail reasoning and decision-making dataset consisting of 14,849 training samples and 1,484 test samples. This dataset takes front-view images as input, and outputs driving decisions, key decision factors, as well as CoT reasoning on the underlying root causes. Figure 8 illustrates the RoadWork scene examples, annotation outputs, and metrics used in this supplementary evaluation.

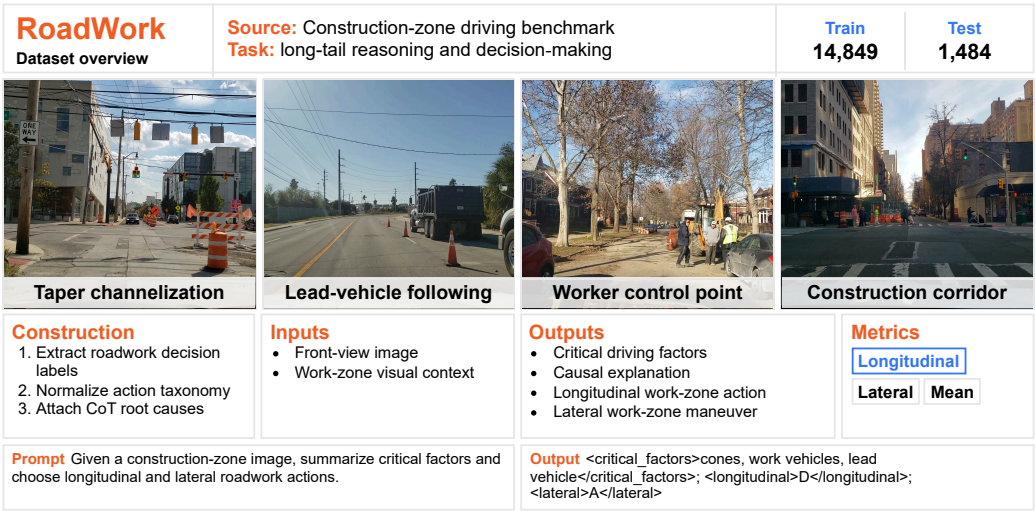


Figure 8: Dataset overview of RoadWork, including representative construction-zone scenes, construction process, input/output fields, and evaluation metrics.

AR1 dataset. Based on the AR1 trajectory prediction data, we further construct a long-tail trajectory prediction dataset for evaluating whether decision-level supervision can benefit downstream planning. Starting from 124,524 raw samples, we first apply a heuristic prefilter that assigns scores according to decision keywords, future-trajectory geometry, and ego-motion history. Specifically, keywords such as pedestrians, riders, construction, cones, obstacles, blocked lanes, and temporary control increase the score; trajectory cues such as large lateral offsets, wide lateral spans, and sign changes indicate non-routine future paths; and motion cues such as hard braking, strong acceleration, and recent speed changes highlight potentially decision-relevant cases. Samples with a heuristic score below 3 are skipped as regular safe scenes, while the remaining candidates are further screened by a vision-language model.

In the VLM screening stage, a sample is retained only when visible evidence supports complex or abnormal long-tail value, including temporary road changes, unusual or oversized objects, vulnerable road user interactions, traffic-control conflicts, severe occlusion, adverse weather, lighting degradation, or multi-factor complexity. The final selected subset contains 15,901 samples, corresponding to a 12.77% selection rate from the raw pool. We split the selected samples into 14,311 training samples and 1,590 test samples using stratified sampling over the screening categories. For AR1, we only construct two SFT formats: an answer-only version that outputs future waypoints in `<answer>` tags, and a decision version that first outputs a driving-decision rationale in `<decision>` tags and then predicts the same waypoint sequence. Each sample takes four sequential front-view images and ego-motion context as input, including historical positions, current speed, and current acceleration. Evaluation follows standard trajectory-prediction metrics, including L2 errors at 1s, 2s, and 3s, as well as ADE and FDE. Figure 9 gives the AR1 overview, including screened long-tail examples and the trajectory-prediction evaluation setup.

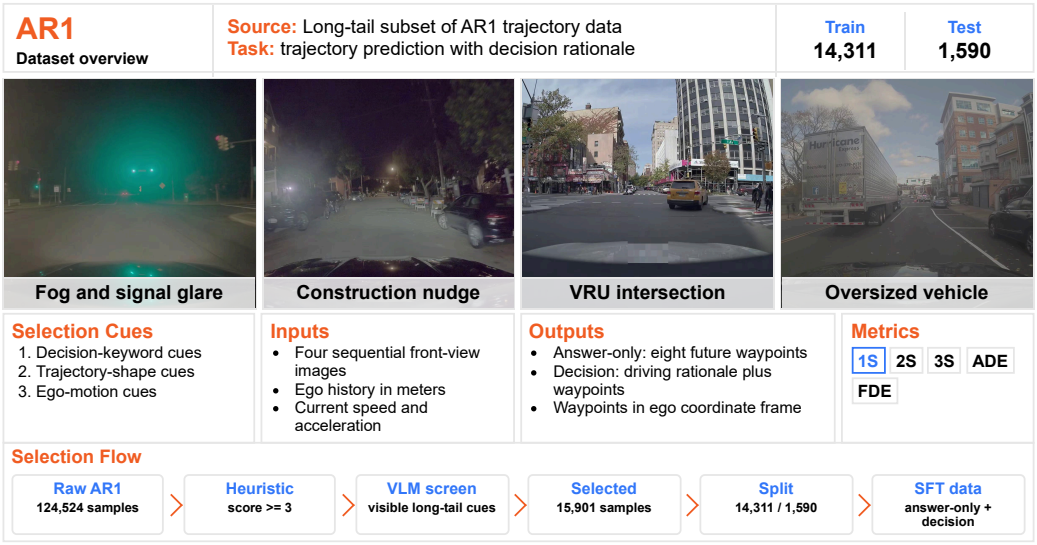


Figure 9: Dataset overview of AR1, including representative screened long-tail scenes, selection rules, input/output fields, and trajectory evaluation metrics.

A.1.2 Experimental Setup and Hyperparameters

To ensure full reproducibility of the CODA-LM training runs, we summarize the hyperparameters of two training stages in Table 4, namely curriculum-guided SFT and chain-of-thought supervised reinforcement learning optimization. The training process of the curriculum-guided SFT stage is divided into two curriculum learning steps. The first step is trained for one epoch, and the second step is trained for three epochs. For CG-SFT, the first step initializes with the pre-trained Qwen3-VL-4B model, and subsequent steps are continually optimized based on the model weights obtained from the previous step. Furthermore, the RSPO algorithm adopts the fully optimized CG-SFT model as its backbone base model for subsequent training.

Table 4: Training hyperparameters on CODA-LM (**left**: CG-SFT; **right**: RSPO).

CG-SFT		**RSPO**	
Hyperparameter	**Value**	**Hyperparameter**	**Value**
Training epochs	(1, 3)	Training epochs	6
Optimizer	AdamW	Actor optimizer	AdamW
Learning rate	4×10^{-5}	Learning rate	2×10^{-6}
Warmup steps	100	Warmup ratio	0.05
Weight decay	0.05	Weight decay	0.1
Train micro-batch size per GPU	6	Gradient clipping	1.0
Gradient accumulation steps	2	Train batch size	64
Global batch size (per optimizer step)*	$12 \times N_{\text{gpu}}$	Micro-batch size per GPU	8
Max sequence length	8192	Clip ratio (low, high)	$(3 \times 10^{-4}, 4 \times 10^{-4})$
Eval micro-batch size per GPU	8	Max prompt length	5120
Eval / save interval (steps)	30	Max response length	4096
Torch dtype	bfloat16	Responses per prompt (n)	4
Trainable modules	ViT, aligner, LLM (full fine-tuning)	Rollout engine / mode	vLLM / async
		VLLM GPU memory utilization	0.8
		VLLM max #seqs	4
		Rollout max sequence length	9216
		Auxiliary SFT loss λ	1×10^{-3}

* Global batch for CG-SFT: micro-batch $\times$ grad. accum. $\times N_{\text{gpu}} = 6 \times 2 \times N_{\text{gpu}}$.

A.1.3 Reinforcement Learning Reward Design

During reinforcement fine-tuning, we use deterministic rule-based reward functions rather than an additional learned reward model. Each rollout is parsed into dataset-specific structured fields, and missing, malformed, or unparsable fields receive zero reward for the corresponding component. The final reward is normalized to $[0, 1]$, which keeps the reward scale consistent across rollouts and makes it directly usable in the group-wise relative advantage computation of RSPO.

For structured decision datasets, the reward is computed as the mean of field-level scores:

$$R(y, y^*) = \frac{1}{|\mathcal{K}|} \sum_{k \in \mathcal{K}} r_k(y, y^*), \tag{10}$$

where y and y^* denote the predicted and ground-truth responses, and $\mathcal{K}$ denotes the set of task-specific fields. For categorical fields, r_k is an exact-match score. For CODA-LM hazardous-source localization, the bounding-box component uses IoU when both prediction and ground truth contain valid boxes in normalized 0–1000 coordinates. If the ground truth explicitly has no hazardous-source box, the prediction receives full box reward only when it explicitly outputs `None`. For CODA-LM longitudinal and lateral suggestions, each field contains two comma-separated labels; the reward is 1.0 when both labels match, 0.5 when one label matches, and 0.0 otherwise. Table 5 summarizes the dataset-specific reward fields and aggregation rules used by RSPO.

Table 5: Rule-based reward definitions used for reinforcement fine-tuning.

Dataset	Reward fields	Field scoring and aggregation
CODA-LM	tail-event flag, scene category, bbox, longitudinal suggestion, lateral suggestion	Exact match for tail event and category; IoU for bbox; two-label partial credit for longitudinal/lateral suggestions; mean over five fields.
Impromptu	VRU state, traffic light, longitudinal decision, lateral decision	Case-normalized exact match for each field; mean over four fields.
RoadWork	longitudinal decision, lateral decision	Exact match for each decision field; mean over two fields.
AR1	eight future waypoints	Parse eight 2D waypoints from the answer; compute ADE; reward is $\max(0, 1 - \text{ADE}/10)$ with a 10m clip scale.

A.1.4 Supplementary Experimental Results

Table 6 provides an additional comparison across different VLM backbones under the same CODA-LM long-tail reasoning evaluation protocol. Although larger or alternative backbones achieve competitive event-detection scores after SFT, Qwen3-VL-4B with curriculum-guided SFT obtains the best overall mean score, mainly due to stronger scene-category recognition, hazardous-source localization, and lateral-decision accuracy.

Table 6: Performance comparison of different vision-language models on the long-tail driving reasoning dataset

Model	Event	Category	Bbox	Long	Lat	Mean
Standard SFT						
InternVL3-8B	82.7	66.8	61.6	61.5	49.9	64.5
MiMo-VL-7B	79.0	61.7	57.0	60.7	53.5	62.4
Qwen3-VL-4B	79.6	63.3	61.1	58.9	54.6	63.5
Qwen2.5-VL-3B	78.4	61.5	54.9	61.8	52.1	61.8
Qwen2.5-VL-7B	77.4	59.3	57.5	59.2	51.6	61.0
Ours (CG-SFT)						
Qwen3-VL-4B	79.2	65.4	62.8	60.2	57.9	65.1

Beyond the aggregate model comparison, we further analyze the two CODA-LM evaluation runs used in the supplementary study. Both runs are evaluated on the same 491-sample test split. Table 7 gives a compact metric summary from standard SFT to curriculum-guided SFT, while Figure 10 visualizes the metric deltas, per-scenario performance, and scene-category confusion matrix.

Table 7: Detailed CODA-LM metric breakdown for the supplementary SFT and CG-SFT evaluation runs. All values are percentages. Tail P/R/F1 are computed from the tail-event flag, with *Yes* as the positive class.

Run	Event	Category	Bbox	Long	Lat	Mean	Tail P	Tail R	Tail F1
SFT	79.6	63.3	61.1	58.9	54.6	63.5	88.9	84.8	86.8
CG-SFT	79.2	65.4	62.8	60.2	57.9	65.1	88.2	85.1	86.6
Δ	-0.4	$+2.0$	$+1.7$	$+1.3$	$+3.4$	$+1.6$	-0.7	$+0.3$	-0.2

The CG-SFT run improves the overall score mainly by strengthening fine-grained category recognition, hazardous-source localization, and the two structured decision fields. Event accuracy is nearly unchanged because the model detects a similar number of hazardous scenes, but it trades a small increase in false positives on safe scenes for slightly higher recall on hazardous scenes. Concretely, the CG-SFT tail-event confusion counts are TP=330, FP=44, FN=58, and TN=59.

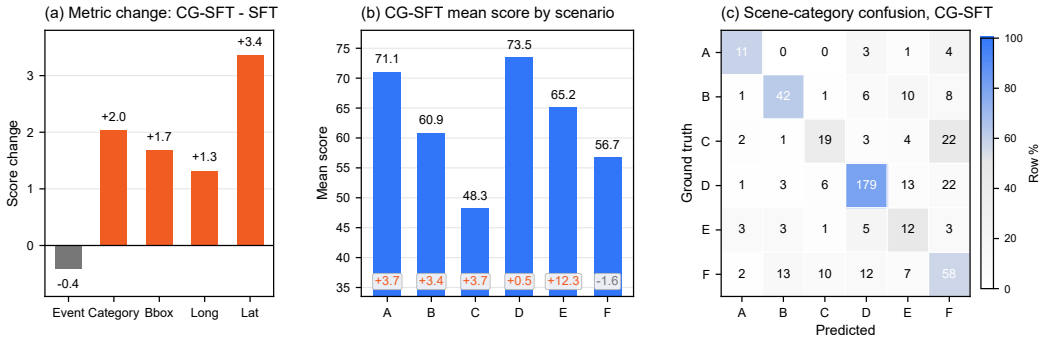


Figure 10: Supplementary CODA-LM statistical analysis. (a) Metric changes from SFT to CG-SFT. (b) CG-SFT mean score by ground-truth scenario category, with bottom annotations showing the change over SFT. (c) CG-SFT scene-category confusion matrix; rows are ground-truth categories and columns are predicted categories. Categories A-F follow the CODA-LM definitions in the dataset description.

The statistical plots reveal two complementary trends. First, the largest scenario-level gain comes from complex combined-risk scenes (E, $+12.3$ mean score), indicating that curriculum-guided supervision helps when multiple visual factors must be jointly associated with the final decision. Second, abnormal traffic participants (C) and construction/rerouting scenes (B) remain difficult: C is often confused with F, while B and E are confused when a combined-risk scene contains construction-like traffic-control cues.

Table 8: Ablation Results for CG-SFT, RSPO, and Their Combination on CODA-LM. CG-SFT denotes curriculum-guided supervised fine-tuning only, RSPO denotes standalone RSPO training, and CG-SFT+RSPO denotes the two-stage pipeline where CG-SFT is performed before RSPO.

Method	Event	Category	Bbox	Long	Lat	Mean
CG-SFT	79.2	65.4	62.8	60.2	57.9	65.1
RSPO	82.9	68.8	64.3	63.9	60.8	68.1
CG-SFT+RSPO	85.7	71.1	64.8	64.5	61.6	69.5

Table 9: Model performance of the RSPO algorithm on the CODA-LM dataset with different values of $\boldsymbol{\beta}$, evaluated from training step 100 to 450.

β	step 100	step 200	step 300	step 350	step 400	step 450
0	66.4	67.8	68.3	69.8	69.1	69.5
0.05	66.1	67.3	67.9	68.4	68.7	68.6
0.1	65.9	67.4	67.7	68.3	68.6	68.6

To further quantify the individual contributions of CG-SFT and RSPO as well as their synergistic effect, we supplement three comparative settings: CG-SFT only, RSPO only, and the combination of CG-SFT and RSPO. The results are summarized in Table 8. We observe that RSPO alone achieves better performance than CG-SFT alone. Integrating CG-SFT and RSPO further improves the model's prediction accuracy and yields the best overall performance. These results verify that CG-SFT and RSPO are highly complementary, and their joint optimization can more effectively boost the overall predictive capability of the model.

Algorithm 1: Curriculum-Guided Supervised Fine-Tuning (CG-SFT)

Input: Curriculum dataset splits $\mathcal{D}_s, \mathcal{D}$;
Pre-trained VLM parameters θ_0;
Optimizer AdamW with learning rate η_s;
Curriculum epochs (E_s, E_f);
Output: Initialized policy parameters θ_{init};

$\theta \leftarrow \theta_0$;
// Curriculum Phase 1: Scene Understanding and Reasoning Traces
for $epoch = 1$ **to** E_s **do**
  for *each batch* $\{(\mathcal{T}_v, \mathcal{T}_l, \mathcal{T}_{y_c})\} \sim \mathcal{D}_s$ **do**
    Compute CoT cross-entropy loss $\mathcal{L}_s(\theta)$ via

$$\mathcal{L}_s(\theta) = -\mathbb{E}_{(\mathcal{T}_v, \mathcal{T}_l, \mathcal{T}_{y_d}, \mathcal{T}_{y_c}) \sim \mathcal{D}} \left[\frac{1}{|\mathcal{T}_{y_c}|} \left(\sum_{m=1}^{|\mathcal{T}_{y_c}|} \log P_\theta(\mathcal{T}_{y_c,m} \mid \mathcal{T}_v, \mathcal{T}_l, \mathcal{T}_{y_c,<m}) \right) \right];$$

    Update θ by gradient descent on $\mathcal{L}_s(\theta)$;
// Curriculum Phase 2: Full Reasoning + Decision
for $epoch = 1$ **to** E_f **do**
  for *each batch* $\{(\mathcal{T}_v, \mathcal{T}_l, \mathcal{T}_{y_c}, \mathcal{T}_{y_d})\} \sim \mathcal{D}$ **do**
    Compute joint CG-SFT loss $\mathcal{L}_s(\theta)$ via Eq. (3);
    Update θ by gradient descent on $\mathcal{L}_s(\theta)$;
$\theta_{\text{init}} \leftarrow \theta$;
return θ_{init};

We conducted an ablation study on the hyperparameter β, with the results presented in Table 9. Under identical training settings, larger values of β lead to slower improvements in reward during training. After training converges, the model with $\beta = 0$ achieves the highest reward. Although some

fluctuations are observed during training, its final reward is substantially higher than those obtained with $\beta > 0$.

A.2 Algorithm Details

A.2.1 Algorithm Instance

To facilitate the understanding of implementation workflows, we provide the detailed pseudocode for both algorithms below. We next detail Algorithm 1 (CG-SFT) and Algorithm 2 (RSPO). Algorithm 1 adopts a two-stage curriculum training paradigm built upon two sub-datasets partitioned from the full dataset. Its first training phase operates on the scene understanding split $\mathcal{D}_s$, where supervision targets only include CoT rationales and structured scene information (Event, Category, Location) for reasoning skill acquisition. The second phase is trained on the full decision dataset $\mathcal{D}$ with comprehensive labels that additionally cover longitudinal and lateral decisions; a joint CG-SFT loss is applied here to align reasoning traces and driving outputs. The final checkpoint obtained from CG-SFT serves as the initial weight initialization for Algorithm 2, namely RSPO. RSPO inherits this pre-trained policy and adopts the same full inference-and-decision dataset for reinforcement optimization. Within each training iteration, the policy samples multiple reasoning-decision rollouts, computes rule-based decision rewards and corresponding advantages, and updates the policy to maximize the decision reward. Meanwhile, a cross-entropy loss regularizes chain-of-thought generation, such that the policy can continuously improve decision quality while preserving the scene reasoning patterns learned in the preceding CG-SFT stage.

Algorithm 2: Reasoning-Supervised Policy Optimization (RSPO)

Input: Dataset $\mathcal{D}$ (CODA-LM);
Initial policy $\pi_{\theta_{\text{init}}}$ from curriculum-guided SFT;
Rule-based reward function R_{rule};
Hyperparameters λ, ϵ, G;
Output: Optimized policy network parameters θ;

$\theta_{\text{old}} \leftarrow \theta_{\text{init}}$;
$\theta \leftarrow \theta_{\text{init}}$;
for *each batch* **do**
  Sample a batch of prompts q and ground-truth responses $(\mathcal{T}^*_{y_c}, \mathcal{T}^*_{y_d})$ from $\mathcal{D}$;
  `// Trajectory Sampling Step:`
  Sample G rollouts $\{(\mathcal{T}_{y_c^i}, \mathcal{T}_{y_d^i})\}_{i=1}^{G} \sim \pi_{\theta_{\text{old}}}(\cdot|q)$;
  Compute decision rewards $r_i = R_{\text{rule}}(\mathcal{T}_{y_d^i}, \mathcal{T}^*_{y_d})$;
  `// Advantage & Ratio Calculation Step:`
  Compute relative advantage $\hat{A}_i$ and sampling ratio $s_i(\theta)$ via Eq. (6);
  Compute the CoT cross-entropy loss $\mathcal{L}_c(\theta)$ on $\mathcal{T}^*_{y_c}$ via Eq. (8);
  `// Policy Update Step:`
  Update θ by maximizing the decision-reward term and minimizing the CoT loss in Eq. (9);
  $\theta_{\text{old}} \leftarrow \theta$;

A.3 Visualization of Experimental Results

The following qualitative examples pair each visualization with the corresponding success and failure analysis.

CODA-LM case analysis. The good cases show that RSPO can connect localized visual hazards to structured driving outputs: it identifies a combined crowd/vehicle risk under reduced visibility, handles an abnormal road-user case where the hazardous participant is relatively small, and remains stable on debris/object and adverse-weather scenes. Across these examples, the predicted category, bounding box, and longitudinal/lateral suggestions remain consistent with the ground truth. The two bad cases reveal a complementary weakness. When the long-tail cue is subtle, peripheral, or visually blended into ordinary traffic infrastructure, the model may classify the scene as safe, suppress the bounding box, and then default to normal driving actions. This suggests that CODA-LM remains challenging for small or off-center safety-critical cues even when the downstream output format is well learned.

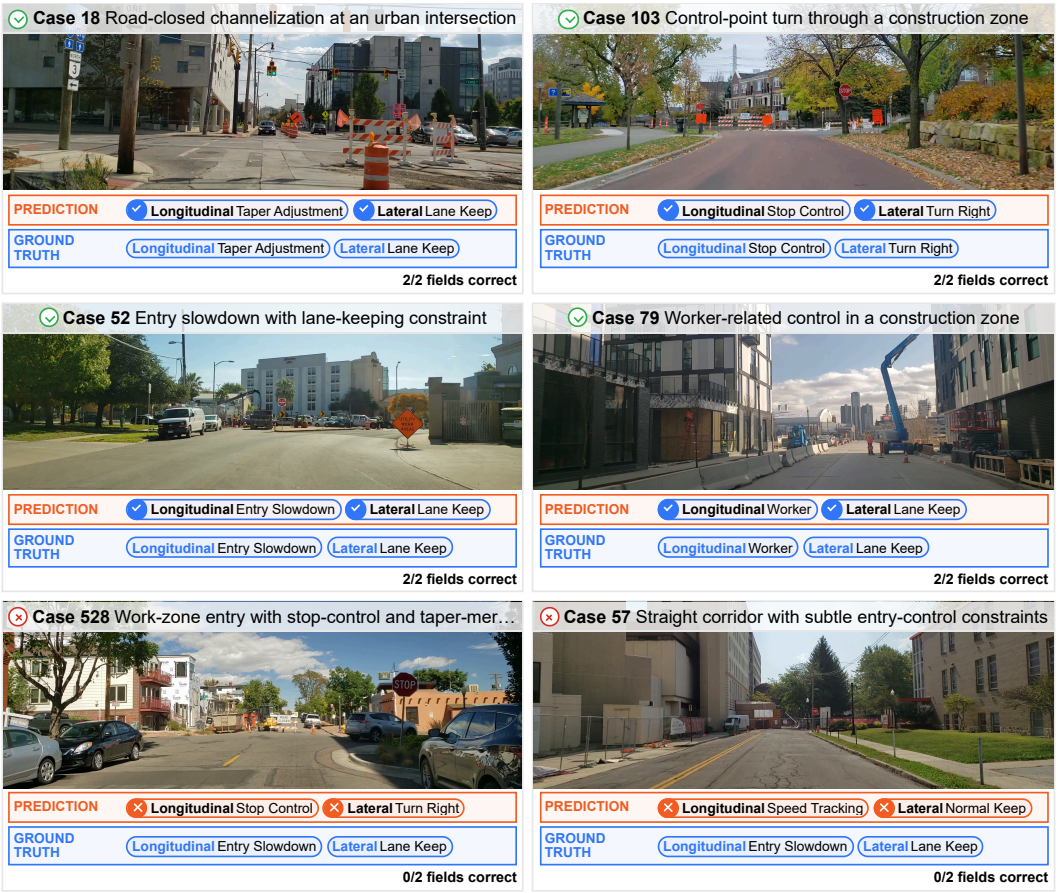


Figure 11: Good and bad RoadWork cases. The top two rows show successful construction-zone reasoning for road-closed channelization, a control-point turn, entry slowdown, and worker-related control. The bottom row shows failures where entry-control, taper-merge, and channelization cues are confused, leading to incorrect longitudinal and lateral actions.

RoadWork case analysis. Figure 11 summarizes the RoadWork examples. The good RoadWork cases show that the model performs well when the construction-zone control semantics are visually explicit: road-closed signs, barrels, barricades, worker cues, and control-point geometry lead to correct taper adjustment, entry slowdown, lane keeping, or turning decisions. The bad cases show that the task becomes harder when several construction cues compete. In one case, the model over-focuses on a stop/control cue and misses the entry-slowdown plus lane-keep label; in another, it treats a constrained work-zone corridor as ordinary speed tracking. These failures illustrate why RoadWork is sensitive to fine-grained semantics: visually nearby cues can imply different longitudinal actions and lateral maneuvers.

Impromptu case analysis. Figure 12 summarizes the Impromptu examples. The good cases highlight different types of consistency. In the roadwork-detour and VRU-aware cases, RSPO recognizes vulnerable-road-user or no-light context and maps the road geometry to the correct turn decision. In the green-light cases, it preserves keep or turn behavior without being distracted by wet roads, roadside objects, or irrelevant visual clutter. The two bad cases are dominated by signal and maneuver ambiguity. The model either misses a red-light/left-turn context or falsely predicts a stop under a green-light approach. These examples show that Impromptu errors are often coupled: a perception error in traffic-light or maneuver context can immediately corrupt the longitudinal or lateral action.

AR1 trajectory case analysis. Figure 13 further visualizes downstream planning behavior on AR1. In the six advantage cases, GRPO often produces over-extended, under-progressed, or nearly straight trajectories, while RSPO stays closer to the ground-truth turning or curved-road path. This pattern

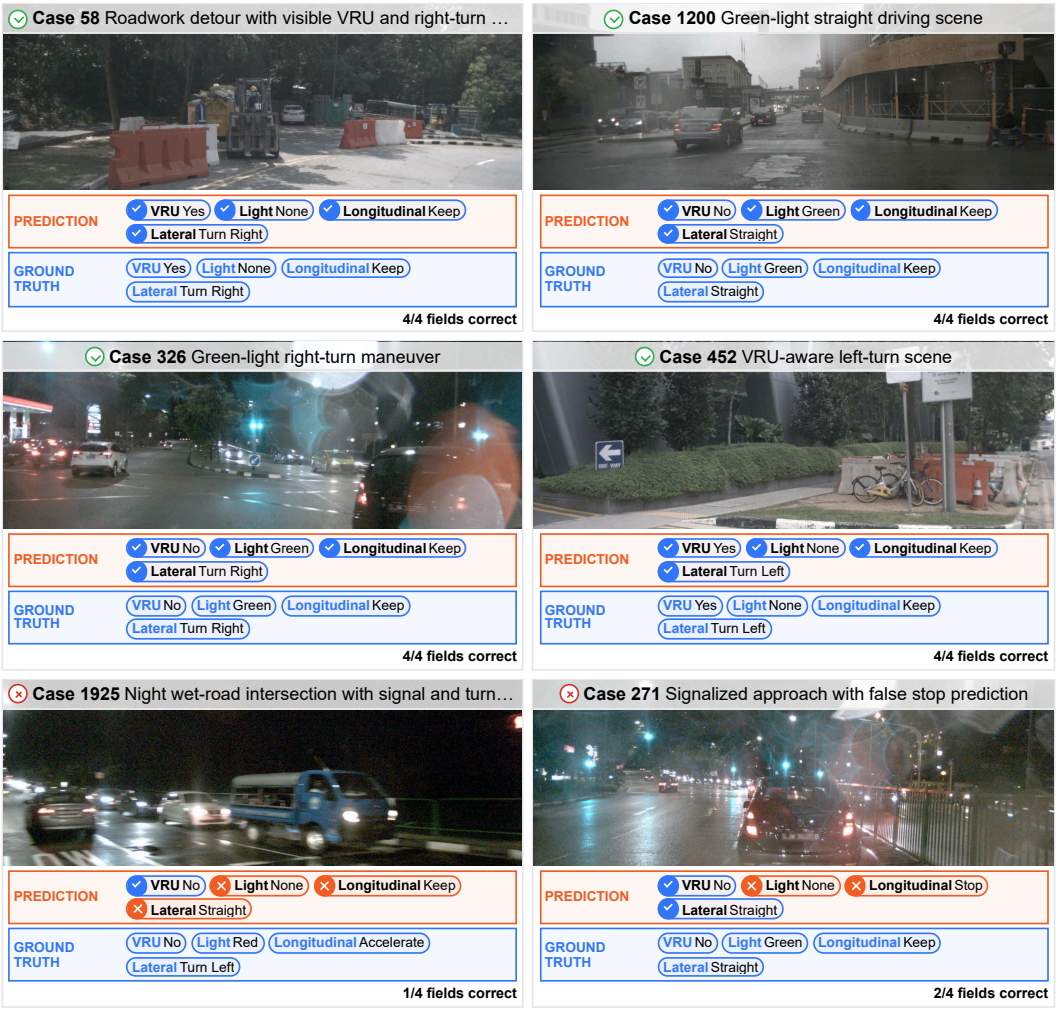


Figure 12: Good and bad Impromptu cases. The top two rows show successful cases covering roadwork detours, VRU reasoning, green-light driving, and turning maneuvers. The bottom row shows failures under night or signalized conditions, where signal perception and maneuver context errors propagate to the longitudinal or lateral decision fields.

suggests that decision-level reasoning supervision helps the model map visual evidence such as wet intersections, road boundaries, night-time curves, and low-visibility highway cues into physically plausible future waypoints. The four failure cases reveal remaining limitations. Under signalized turning, vulnerable-road-user interactions, narrow industrial lanes, or ambiguous residential bends, RSPO can still cut inside the intended path or under-estimate the turning radius, indicating that long-tail planning remains sensitive to fine-grained lane geometry and maneuver cues.

Overall qualitative analysis. Across the four datasets, successful cases share a common pattern: the model grounds the safety-critical visual evidence and maps it to a consistent structured decision or trajectory. The failure modes differ by dataset: CODA-LM errors are often caused by small or peripheral hazards, Impromptu failures are usually tied to signal and maneuver ambiguity, RoadWork errors come from fine-grained construction-zone semantics, and AR1 trajectory errors arise when lane geometry or temporary control cues are difficult to infer. These examples support the motivation for RSPO: decision rewards and supervised reasoning consistency should be optimized together so that visual evidence, causal interpretation, and final actions remain aligned.

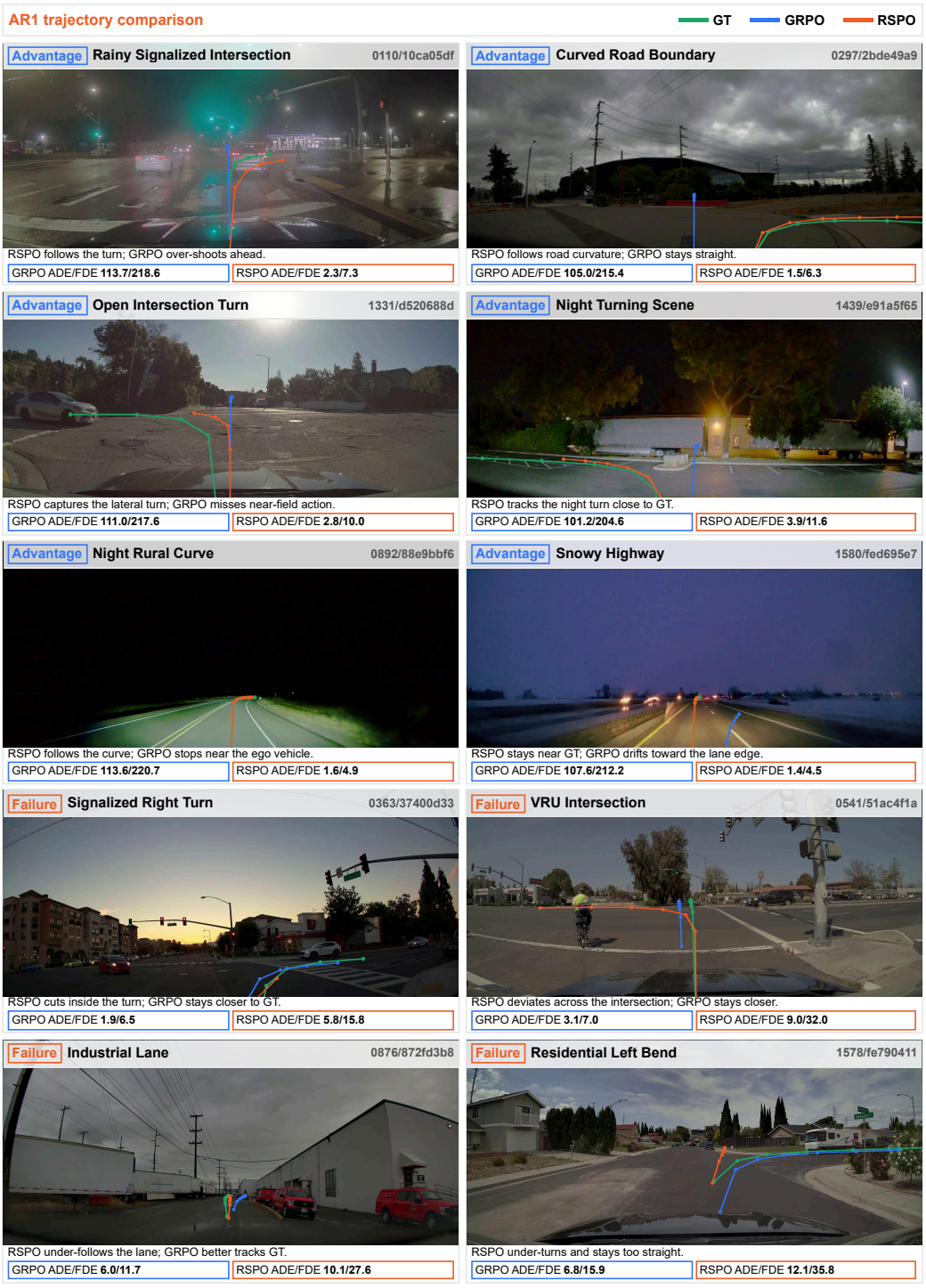


Figure 13: Qualitative AR1 trajectory comparison between GT, GRPO, and RSPO. The top six cases are representative advantage cases where RSPO better follows the ground-truth path under rain, curved-road geometry, intersection turning, night-time turning, rural curves, and snowy highways. The bottom four cases show failure modes where RSPO cuts turns too tightly, deviates around vulnerable-road-user interactions, under-follows narrow lanes, or under-turns on a residential bend. ADE/FDE values are computed over the eight predicted future waypoints.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and Section 1 state the paper's scope, contributions, and main empirical claims. These claims are matched by the methodology in Section 3, the experiments in Section 5, and the limitations discussed in the conclusion.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The conclusion includes a dedicated limitations paragraph discussing the offline-dataset evaluation scope, the lack of closed-loop or real-world validation, and the dependence on structured annotations and rule-based rewards.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [N/A]

Justification: The paper does not present theorem-style guarantees or formal proof claims. The assumptions and optimization objectives used by CG-SFT and RSPO are stated in Section 3.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: Section 5 and the appendix describe the datasets, model backbones, baselines, evaluation metrics, reward functions, and training hyperparameters needed to reproduce the main comparisons. The appendix also gives dataset construction details, train/test splits, and the RSPO optimization loop.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: we submit the code in supplemental material.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Section 5 specifies the datasets, baselines, evaluation metrics, and downstream planning setup. The appendix provides dataset sizes and splits, reward definitions, optimizer choices, learning rates, batch sizes, sequence lengths, and other training hyperparameters.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: The paper reports the main comparisons over multiple random seeds, with all runs evaluated on the same held-out test splits. We summarize seed-level variation using mean performance and standard deviation.

   Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The detail information is add in the supplemental data.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The work uses existing autonomous-driving benchmarks and derived annotations for offline research, and it does not involve deployment on public roads or interaction with human subjects. We have reviewed the NeurIPS Code of Ethics and are not aware of deviations from it.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [N/A]

Justification: This work without any negative social impacts.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: The paper does not release a new high-risk pretrained model, image generator, or scraped internet dataset. The experiments are conducted on offline autonomous-driving benchmark data and derived structured annotations.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [No]

Justification: The paper cites the datasets, model backbones, and baseline methods used in the experiments, including CODA-LM, Impromptu, RoadWork, AR1, and Qwen-family models. However, the current manuscript does not explicitly enumerate the licenses and terms of use for each existing asset.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: The paper introduces derived long-tail reasoning-and-decision annotations and documents their construction pipelines, input/output fields, label definitions, train/test splits, and evaluation protocols in Section 5 and the appendix. The assets themselves are not yet openly released, as noted above.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: The paper does not conduct crowdsourcing experiments or human-subject studies. Manual verification is used for annotation quality control, but no participant-facing experiment is performed.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.

- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The research does not involve human subjects or participant studies, so IRB approval or equivalent review for human-subject research is not applicable.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: LLMs were used only for writing, editing, and language polishing. This use did not affect the core methodology, experimental design, scientific rigor, or originality of the research.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.