DRIVESCOUT: SHARPENING MLLM PERCEPTION IN AUTONOMOUS DRIVING WITH ADAPTIVE AND ROBUST VISUAL FOCUSING

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal Large Language Models (MLLMs) are increasingly used in autonomous driving for perception, reasoning, and interpretable scene understanding. Yet safety-critical small-scale targets, such as pedestrians, traffic signs, and distant vehicles, remain difficult to recognize, particularly when motion blur weakens their already limited visual evidence. Think-with-Images MLLMs can acquire additional evidence by cropping and enlarging local regions, but blur may destabilize the selected region, while indiscriminate image operations add inference cost. To study this problem, we introduce SCOUT, a paired driving VQA benchmark built from driving scenes in nuScenes, where each clear observation is matched with motion-degraded variants for small-target understanding. We further propose DriveScout, a reinforcement learning framework that formulates visual focusing as selective evidence acquisition. DriveScout combines Target-Grounded Spatial Credit with Necessity-Aware Focus Optimization, while a Progressive Blur Curriculum supports learning as the visual evidence becomes increasingly degraded. Experiments on SCOUT and two additional driving benchmarks show that DriveScout improves performance under matched motion-degradation settings while reducing redundant visual operations.

1 INTRODUCTION

As Multimodal Large Language Models (MLLMs) (Bai et al., 2025; Yang et al., 2025a; Chen et al., 2025; Liu et al., 2023; Guo et al., 2025) continue to advance, their visual question answering (VQA) and language-based scene understanding capabilities have improved substantially. MLLMs therefore hold strong potential for autonomous driving (Li et al., 2026b; Liu et al., 2025; Zeng et al., 2025; Mao et al., 2023b), providing a unified framework for perception, reasoning, and explainable decision-making (Sima et al., 2024; Chen et al., 2024; Mao et al., 2023a).

Existing MLLM-based studies in autonomous driving often formulate scene understanding, reasoning, and explanation as VQA-style perception tasks (Ma et al., 2024; Corbiere et al., 2025; Guo et al., 2024). However, they largely overlook safety-critical small-scale targets: as illustrated in Figure 1, distant vehicles, pedestrians, and traffic signs are often small and difficult to recognize, yet provide

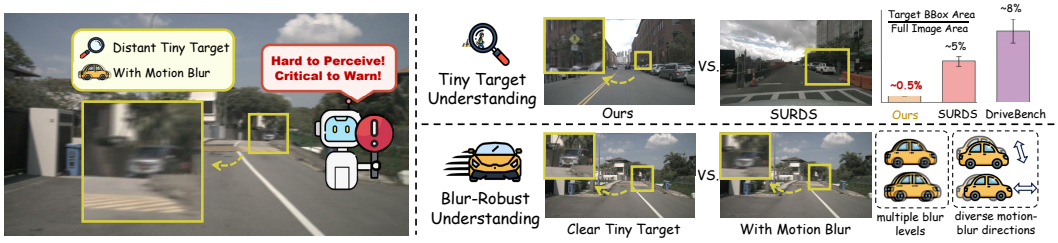


Figure 1: **Motivation and overview of SCOUT.** Small-scale safety-critical targets are easily overlooked, particularly under motion blur. SCOUT provides driving VQA samples for distant targets with paired clear and motion-degraded observations.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

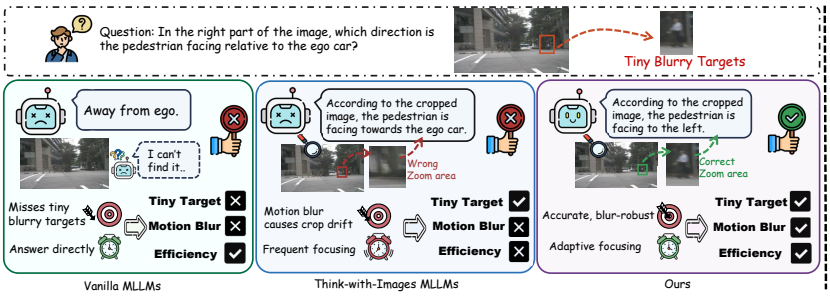


Figure 2(a). Comparing ours with vanilla MLLMs and Think-with-Images MLLMs

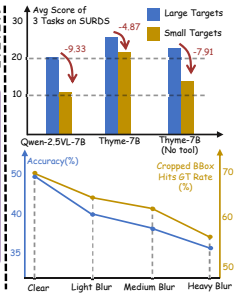


Figure 2(b). Motivation Experiments

Figure 2: **Motivation of DriveScout.** (a) Existing MLLMs struggle with tiny blurry targets, crop drift, and redundant operations; DriveScout enables robust adaptive focusing. (b) Think-with-Images improves small-target understanding, but blur reduces crop accuracy and robustness.

important cues for downstream driving decisions. This raises a key question: *how can MLLMs effectively understand such targets?*

Inspired by Think-with-Images MLLMs (OpenAI, 2025), we introduce visual focusing into autonomous driving for small-scale target understanding. By focusing on informative regions through operations such as cropping and zooming, the model acquires more discriminative evidence for distant targets that are difficult to recognize from the original image alone (Zhang et al., 2025c; Zheng et al., 2025b; Yang et al., 2025b; Zhang et al., 2025b). As shown in the upper part of Figure 2(b), local focusing substantially improves small-scale target understanding.

However, motion blur in autonomous-driving scenarios (Xie et al., 2025; Liao et al., 2026) challenges the Think-with-Images paradigm. As shown in the lower part of Figure 2(b), stronger blur reduces the cropping accuracy of typical Think-with-Images models and degrades performance. Blur corrupts scarce local cues and destabilizes target localization, causing focus-region crops to drift from the true target and weakening local evidence acquisition. This raises our first challenge: **I)** *how can we enable MLLMs to acquire reliable local evidence for small-scale target understanding under motion-degraded observations?*

To study this challenge, we introduce SCOUT, a dedicated paired QA benchmark for matched evaluation of distant small-target understanding under parameterized motion degradation. Existing autonomous-driving datasets (Ishaq et al., 2025; Li et al., 2024; Qian et al., 2024; Kim et al., 2018) mainly focus on scene-level understanding or salient entities, with limited attention to distant small targets. SCOUT covers the safety-critical tasks of vehicle orientation, pedestrian motion, and traffic sign understanding. Each clear observation is paired with motion-degraded variants spanning different severities and directions, enabling factor-isolated evaluation while preserving the scene, target, and question.

Moreover, image operations in the Think-with-Images paradigm usually require additional computation, including image cropping, zooming, and subsequent model inference over processed visual inputs(Zhang et al., 2025a). In autonomous driving, such extra computation is particularly undesirable(Tian et al., 2024; Rawal et al., 2026). For large and clearly visible targets, which are often closer to the ego vehicle and require more timely responses, as well as for small-scale targets that are already sufficiently clear in the original image, image operation is often unnecessary and offers only marginal gains(Zhang et al., 2025a). This raises our second challenge: **II)** *how can we guide MLLMs to adaptively focus on local regions only when necessary, balancing recognition accuracy and inference cost?*

To address these challenges, we propose DriveScout, a reinforcement learning framework that treats adaptive visual focusing as *selective evidence acquisition*. For Challenge I, Target-Grounded Spatial Credit (TGSC) and a Progressive Blur Curriculum (PBC) jointly teach where to acquire reliable local evidence under progressively degraded observations. For Challenge II, Necessity-Aware Focus Optimization (NAFO) uses a counterfactual direct-view response to distinguish necessary from redundant focusing, rewarding useful inspection while suppressing crop supervision when the original view is sufficient. Together, these components teach the policy both *where* to focus reliably under blur and *when* the additional inference is warranted.

Our main contributions are summarized as follows:

1. We identify and formulate **safety-critical small-scale target understanding** in autonomous driving, highlighting distant vehicles, pedestrians, and traffic signs whose evidence is easily weakened by motion blur during safety-critical decisions.
2. We introduce **SCOUT**, a paired driving VQA benchmark for safety-critical small-scale targets, covering vehicle orientation, pedestrian motion, and traffic sign understanding with matched motion-degraded variants for reliable robustness evaluation.
3. We propose **DriveScout**, which casts visual focusing as selective evidence acquisition and combines TGSC, PBC, and NAFO. The unified design encourages reliable local evidence acquisition under blur while reducing redundant image operations when the original view is sufficient.

2 Related Works

2.1 Vision-Language Models for Autonomous Driving

Vision-language models have been increasingly applied to autonomous driving for perception, reasoning, planning, and decision making (Choudhary et al., 2024; Fu et al., 2024; Hwang et al., 2024; Jiang et al., 2024; Liao et al., 2025; Shao et al., 2024; Li et al., 2026a; Zhang et al., 2026; Chen et al., 2026). Recent driving VQA and reasoning benchmarks formulate scene understanding as language-based reasoning or question answering (Qian et al., 2024; Sima et al., 2024; Nie et al., 2024; Ishaq et al., 2025; Tao et al., 2026), while DriveBench (Xie et al., 2025) evaluates MLLMs on constructed motion- and zoom-blur corruptions. These works mainly emphasize scene understanding, planning, or salient traffic participants (Cao et al., 2024; Li et al., 2024; Marcu et al., 2024; Yu et al., 2025), whereas we target distant small-scale safety-critical objects under motion-blur corruptions.

2.2 Small-Scale Object Recognition and Visual Degradation

Small-scale target perception is important in driving scenes and has been studied in detection-oriented datasets and benchmarks (Cheng et al., 2023; Yu et al., 2020). Robustness and corruption benchmarks further evaluate model reliability under adverse visual conditions, but mostly focus on detection, segmentation, or general perception (Sakaridis et al., 2021; Bijelic et al., 2020; Dong et al., 2023; Qiao et al., 2026). In contrast, our dataset studies QA-style reasoning over small-scale driving objects and explicitly evaluates sensitivity to matched motion-degraded inputs.

2.3 Tool-Augmented Visual Reasoning for Autonomous Driving

Recent MLLMs perform active visual reasoning by cropping, zooming, segmenting, or manipulating visual inputs to gather additional evidence (Zhang et al., 2025c;a; Zheng et al., 2025b). Driving-oriented agentic and tool-augmented methods also show the potential of active perception in autonomous driving (Qian et al., 2025; Zheng et al., 2025a; Luo et al., 2025; Peng et al., 2026; Bassole et al., 2026). However, they rarely consider whether visual operations are necessary, how blur affects crop accuracy, or how to train models to adapt to blur progressively. We address these issues through a Progressive Blur Curriculum, Target-Grounded Spatial Credit, and Necessity-Aware Focus Optimization.

3 SCOUT Dataset

Small targets in driving scenes provide important cues for downstream driving decisions, such as anticipating the motion of distant traffic participants and recognizing traffic signs in advance. However, existing autonomous driving VQA datasets or benchmarks often focus on holistic scene understanding or salient objects, leaving small-target understanding underexplored. (Marcu et al., 2024; Cao et al., 2024) This issue becomes more challenging under motion blur, where small objects easily lose visual details. To address this gap, we introduce SCOUT, a paired small-target VQA benchmark for autonomous driving that focuses on safety-critical traffic participants and traffic signs. SCOUT pairs each clear observation with motion-degraded variants spanning different severities and directions, supporting matched evaluation under degraded visual evidence.

3.1 Dataset Construction

Small-target candidate selection. SCOUT is built from nuScenes(Caesar et al., 2020) camera images and 3D annotations. We first project the 3D annotations onto the corresponding camera views and compute 2D bounding boxes. To ensure both visibility and small-target difficulty, we retain targets that satisfy predefined constraints on visibility, in-image coverage, occlusion, and image-area ratio, while removing heavily truncated or occluded targets.

Question-answer construction. For each retained target, we construct a task-specific QA pair. Vehicle-heading and pedestrian-motion answers are derived from the corresponding nuScenes annotations, whereas traffic-sign semantics are manually annotated from the source images. The resulting samples cover three complementary driving tasks: vehicle heading, pedestrian motion, and traffic sign understanding for reliable target-centric evaluation.

Paired motion-degradation construction. Each source image is retained as the clear reference and paired with matched motion-degraded variants spanning multiple severities and directions. Because each pair preserves the scene, target, question, and answer, performance differences isolate sensitivity to applied degradation from semantic changes.

Human verification. We manually verify every candidate QA instance against the source image, an enlarged target crop, and every degraded version. An instance is retained only if the target is visible, its answer remains visually verifiable in every version, no other plausible target causes ambiguity, and the enlarged crop contains sufficient context. After verification, SCOUT contains 4,769 QA pairs, with 1,000 used for testing.

Scene-disjoint split. We split SCOUT by nuScenes scene identifiers, assigning all QA pairs from each scene exclusively to either the training or test set. Thus, no source scene or camera observation is shared between the two splits, preventing same-scene leakage.

3.2 Reasoning Trajectory Construction for SFT

Beyond QA annotations, we construct reasoning trajectories for the training split to support SFT. For samples requiring local inspection, we use GPT-5 to generate image-operation trajectories with code templates, where the enlarged bounding box serves as the crop region. For easy or boundary samples, we generate trajectories without image operations. These trajectories are used only as SFT training supervision and teach the model to use visual focusing only when needed.

4 DriveScout: Selective Visual Evidence Acquisition

4.1 Problem Formulation

Given an image-question pair $x = (I, q)$, a Think-with-Images MLLM may either answer directly from the original image or invoke a crop-based focus operation before producing its answer. We view this behavior as *selective visual evidence acquisition*: the model must determine whether additional local evidence is needed and, when focusing is used, which region should provide that evidence. A useful acquisition policy should therefore satisfy two complementary requirements. It should produce spatially grounded crops under degraded observations, and it should avoid focus operations when the original view already contains sufficient evidence.

These requirements correspond directly to our two challenges. DriveScout addresses Challenge I through Target-Grounded Spatial Credit (TGSC) and a Progressive Blur Curriculum (PBC), and Challenge II through Necessity-Aware Focus Optimization (NAFO). A necessity gate connects where the model focuses to whether the added evidence is useful.

We optimize this policy with Group Relative Policy Optimization (GRPO) (Guo et al., 2025). Given x, the old policy $\pi_{\theta_{\text{old}}}$ samples a group of G response trajectories $\{y_i\}_{i=1}^{G}$ with rewards $\{r_i\}_{i=1}^{G}$. The group-normalized advantage is $A_i = (r_i - \mu_r)/(\sigma_r + \epsilon)$, where μ_r and σ_r are the reward mean and standard deviation within the group. The policy is optimized with the clipped objective

$$\mathcal{L}_{\text{GRPO}}(\theta) = -\frac{1}{G}\sum_{i=1}^{G} \min\big(\rho_i A_i,\ \text{clip}(\rho_i, 1-\eta, 1+\eta)A_i\big) + \beta D_{\text{KL}}\big(\pi_\theta(\cdot|x)\,\|\,\pi_{\text{ref}}(\cdot|x)\big), \quad (1)$$

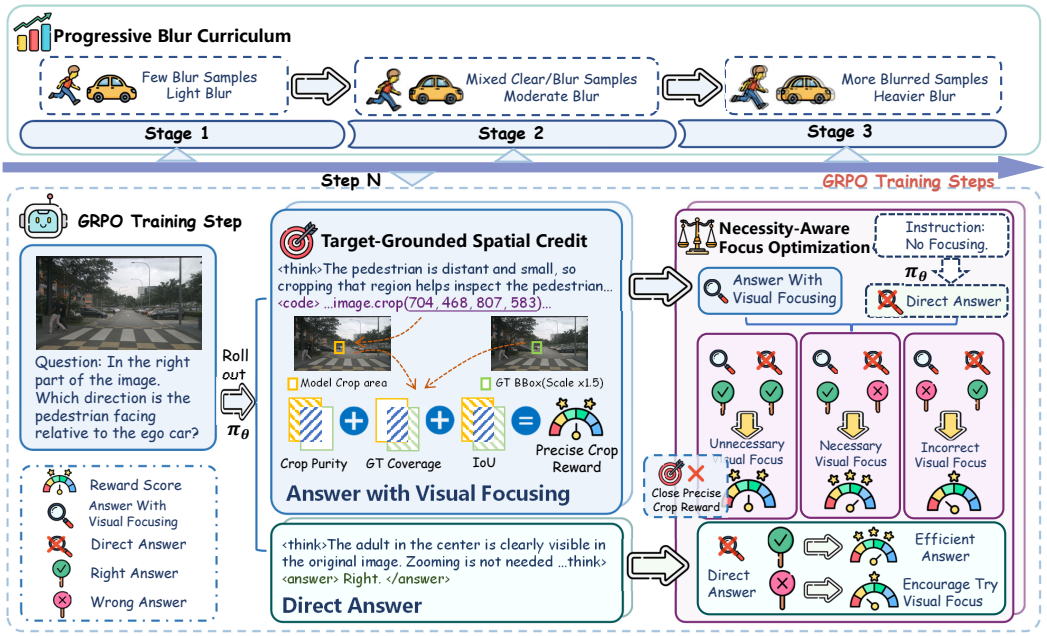


Figure 3: **Overview of DriveScout.** DriveScout combines Target-Grounded Spatial Credit and a Progressive Blur Curriculum for reliable evidence acquisition under blur (Challenge I), together with Necessity-Aware Focus Optimization for selective focusing (Challenge II).

where $\rho_i = \pi_\theta(y_i|x)/\pi_{\theta_{\text{old}}}(y_i|x)$ is the policy probability ratio, η is the clipping coefficient, and β controls the strength of KL regularization. DriveScout augments this outcome-level objective with signals for evidence necessity and crop quality, as summarized in Figure 3.

4.2 Necessity-Aware Focus Optimization

Counterfactual evidence necessity. An observed focus operation does not by itself indicate that the local view was useful: the same question may already be answerable from the original image. We therefore estimate evidence necessity with a counterfactual direct-view response. Let $f_i \in \{0, 1\}$ indicate whether rollout i invokes visual focusing, and let $a_i^{\text{main}} \in \{0, 1\}$ indicate whether its final answer is correct. For each correct focus-assisted rollout, we query the same policy again with the same image and question while disallowing image operations. The resulting response is used only for necessity estimation and is not included as an additional policy-optimization rollout.

Let $a_i^{\text{dv}} \in \{0, 1\}$ denote whether this counterfactual direct-view response is correct. For each eligible focused rollout, we define the operational necessity signal as

$$\nu_i = 1 - a_i^{\text{dv}}, \qquad f_i = 1,\ a_i^{\text{main}} = 1. \tag{2}$$

Thus, $\nu_i = 1$ means that focusing succeeds while the direct-view response fails, indicating useful local evidence; $\nu_i = 0$ marks a correct but redundant focus operation. For ineligible rollouts, we set $a_i^{\text{dv}} = 0$ by convention to maintain consistent reward computation during training.

Within NAFO, we incorporate this signal through a necessity-aware focus reward:

$$r_i^{\text{focus}} = \lambda_{\text{dir}}(1 - f_i)a_i^{\text{main}} + \lambda_{\text{nec}} f_i a_i^{\text{main}}(1 - a_i^{\text{dv}}) - \lambda_{\text{red}} f_i a_i^{\text{main}} a_i^{\text{dv}} - \lambda_{\text{fail}} f_i (1 - a_i^{\text{main}}). \tag{3}$$

The four terms respectively reward a correct direct answer, reward a counterfactually necessary focus operation, penalize a correct but redundant one, and penalize an unsuccessful focus-assisted attempt. This directly addresses Challenge II by encouraging the crop-and-inference path only when local inspection changes the evidence available for answering.

4.3 Target-Grounded Spatial Credit

Target-grounded crop quality. When local evidence is acquired, its spatial quality is equally important. A focusing crop should not merely reproduce a tight detection box: it should cover

the target, retain useful surrounding context, and exclude excessive irrelevant content. Given a $1.5\times$-enlarged ground-truth box B_g and a predicted crop B_p parsed from the generated operation, let $A_\cap = |B_p \cap B_g|$. We then compute three complementary geometric measures:

$$\mathrm{Cov}_{gt} = \frac{A_\cap}{|B_g|}, \quad \mathrm{Pur}_{crop} = \frac{A_\cap}{|B_p|}, \quad \mathrm{IoU} = \frac{A_\cap}{|B_p \cup B_g|}. \tag{4}$$

GT Coverage measures whether the crop contains the target, Crop Purity discourages overly loose regions, and IoU provides a complementary overlap measure. Their weighted combination is

$$s_{\mathrm{crop}} = w_{\mathrm{cov}}\mathrm{Cov}_{gt}(B_p, B_g) + w_{\mathrm{pur}}\mathrm{Pur}_{crop}(B_p, B_g) + w_{\mathrm{iou}}\mathrm{IoU}(B_p, B_g), \tag{5}$$

where w_{cov}, w_{pur}, and w_{iou} are weighting coefficients. We additionally penalize crops that cover too little of the target or occupy an excessive fraction of the original image. Let $\mathbb{I}_c = \mathbb{I}[\mathrm{Cov}_{gt}(B_p, B_g) < \tau_c]$ and $\mathbb{I}_a = \mathbb{I}[|B_p|/|\Omega| > \tau_a]$. The resulting reward is

$$r_{\mathrm{crop}} = s_{\mathrm{crop}} - \lambda_{\mathrm{cov}}\mathbb{I}_c - \lambda_{\mathrm{area}}\mathbb{I}_a, \tag{6}$$

where Ω is the original image region, and τ_c and τ_a are the coverage and area thresholds.

Necessity-gated spatial credit assignment. Applying r_{crop} to the entire completion would mix spatial supervision with unrelated reasoning and answer tokens. For rollout i, we therefore use a binary mask $M_{i,t}^{\mathrm{crop}} \in \{0, 1\}$ to identify coordinate-generation tokens, compute a crop advantage A_i^{crop} within its GRPO group, and gate this spatial signal with the counterfactual necessity estimate:

$$g_i^{\mathrm{nec}} = 1 - f_i a_i^{\mathrm{main}} a_i^{\mathrm{dv}}, \qquad \widetilde{A}_{i,t}^{\mathrm{crop}} = g_i^{\mathrm{nec}} M_{i,t}^{\mathrm{crop}} A_i^{\mathrm{crop}}. \tag{7}$$

The token mask localizes credit to the coordinates that determine the crop, while g_i^{nec} removes spatial supervision from focus operations verified to be redundant. In particular,

$$f_i = 1,\ a_i^{\mathrm{main}} = 1,\ a_i^{\mathrm{dv}} = 1 \quad \Rightarrow \quad g_i^{\mathrm{nec}} = 0 \quad \Rightarrow \quad \widetilde{A}_{i,t}^{\mathrm{crop}} = 0,\ \forall t. \tag{8}$$

The resulting crop-specific objective $\mathcal{L}_{\mathrm{crop}}$ uses $\widetilde{A}_{i,t}^{\mathrm{crop}}$ to optimize crop-coordinate tokens. Target-grounded evaluation and token-localized credit provide reliable spatial supervision for Challenge I under motion blur, while the gate couples it to Challenge II.

4.4 Progressive Blur Curriculum

Strong blur weakens target cues and can make answer and crop signals unreliable early in training. For Challenge I, we therefore expose the policy to increasingly difficult observations using a Progressive Blur Curriculum. Each sample has one clear image and three motion-blurred variants:

$$v \in \{\mathrm{clear}, \mathrm{blur}_{\ell_1}, \mathrm{blur}_{\ell_2}, \mathrm{blur}_{\ell_3}\}. \tag{9}$$

All variants remain in the mixture, but sampling progressively shifts toward stronger blur:

$$P_s = \{\mathrm{clear} : q_s^0, \mathrm{blur}_{\ell_1} : q_s^1, \mathrm{blur}_{\ell_2} : q_s^2, \mathrm{blur}_{\ell_3} : q_s^3\}, \quad s \in \{1, 2, 3\}. \tag{10}$$

Here $q_s^0, q_s^1, q_s^2, q_s^3$ are the stage-wise probabilities for clear inputs and light, medium, and strong blur. Early stages establish answer and localization behavior from clearer evidence; later stages emphasize degraded observations to learn useful crops under blur.

4.5 Unified Training Objective

The outcome-level reward combines answer correctness, response format, and evidence necessity:

$$r_i = r_i^{\mathrm{ans}} + \lambda_{\mathrm{fmt}} r_i^{\mathrm{fmt}} + \lambda_{\mathrm{focus}} r_i^{\mathrm{focus}}, \tag{11}$$

where r_i^{ans} evaluates answer correctness, r_i^{fmt} rewards valid formatting, and r_i^{focus} denotes the necessity-aware focus reward used by NAFO in Eq. 3.

We use the combined reward r_i to compute the group-normalized advantage in the GRPO objective. In parallel, the crop-specific objective $\mathcal{L}_{\mathrm{crop}}$ applies the necessity-gated spatial advantage in Eq. 7 only to crop-coordinate tokens.

The final training objective is:

$$\mathcal{L} = \mathcal{L}_{\mathrm{GRPO}} + \lambda_{\mathrm{crop}}\mathcal{L}_{\mathrm{crop}}. \tag{12}$$

where λ_{crop} controls spatial supervision. During training, PBC controls the visual-variant sampling distribution, TGSC supplies crop-coordinate credit, and NAFO regulates whether local focusing should be reinforced. The objective jointly learns the two decisions posed by our challenges: where to acquire reliable local evidence under blur, and whether it is necessary for the current sample.

Table 1: Main results on SCOUT under clear and blurred settings. Ped., Veh., and Sign denote pedestrian, vehicle, and traffic-sign questions. $C-B$ denotes Clear − Blur, and Avg. Drop is Clear Avg. − Blur Avg.; lower values indicate greater robustness to blur. Best and second-best results among open-source and Think-with-Images MLLMs are bolded and underlined; closed-source models are reported for reference.

Method	Blur				Avg. Drop↓	Clear			
	Ped. ($C-B$)	Veh. ($C-B$)	Sign ($C-B$)	Avg.		Ped.	Veh.	Sign	Avg.
Closed-source MLLMs									
Doubao-Seed-1.6	34.54 (14.43)	45.33 (11.62)	64.74 (12.72)	44.50	12.90	48.97	56.95	77.46	57.40
GPT-4.1	43.30 (6.18)	61.28 (1.82)	61.85 (20.23)	54.40	6.70	49.48	63.10	82.08	61.10
Open-source MLLMs									
Qwen2.5VL-32B-Instruct	38.95 (5.64)	32.47 (5.57)	49.71 (12.14)	38.30	6.40	44.59	38.04	61.85	44.70
Qwen3VL-8B-Instruct	45.62 (7.47)	41.91 **(0.23)**	53.75 (11.57)	45.40	5.00	53.09	42.14	**65.32**	50.40
InternVL2.5-8B	32.47 (7.22)	37.81 (2.74)	31.79 (17.92)	34.70	7.10	39.69	40.55	49.71	41.80
Qwen2.5VL-7B-Instruct	34.02 (7.73)	37.36 (3.19)	38.15 (12.14)	36.20	6.50	41.75	40.55	50.29	42.70
Think-with-Images MLLMs									
Thyme-7B-RL	37.37 (8.25)	33.49 (3.18)	52.02 (11.56)	38.20	6.60	45.62	36.67	63.58	44.80
Base (SFT + GRPO)	45.88 (6.96)	51.25 (6.84)	48.55 (14.46)	48.70	8.20	52.84	58.09	63.01	56.90
DriveScout	**48.97 (5.15)**	**58.31** (0.69)	**55.49 (5.20)**	**54.20**	**3.20**	**54.12**	**59.00**	60.69	**57.40**

5 Experiments

5.1 Experimental Setup

Evaluation metrics. For SCOUT, we use accuracy (Acc.) as the primary evaluation metric. To measure sensitivity to the applied motion-degradation protocol, we report the performance gap between matched clear and motion-degraded observations, denoted as *Clear-Blur*. Each pair shares the same scene, target, question, and answer, so the gap measures how model performance changes as the available visual evidence is degraded. A smaller *Clear-Blur* gap indicates better performance preservation under the applied degradation.

For SURDS(Guo et al., 2024), we follow the official evaluation protocol and its task-specific scoring functions. For DriveBench(Xie et al., 2025), we adopt the official GPT-based metric, GPT-Score, to evaluate response correctness.

Implementation Details. We use Thyme-RL-7B(Zhang et al., 2025c) as the base model. To adapt the model to our driving scenarios and task format, we first conduct supervised fine-tuning (SFT) for one epoch with a learning rate of 1×10^{-6}. For reinforcement learning, we train the model with GRPO for five epochs, where each prompt is sampled with eight responses. All experiments are conducted on 8 NVIDIA H20 GPUs, each with 96 GB of memory.

5.2 Main Results

5.2.1 Performance on SCOUT

We first evaluate our method on SCOUT and compare it with a broad set of open-source and closed-source state-of-the-art MLLMs. The results are shown in Table 1. We make two main observations.

DriveScout achieves strong performance under blurred small-object QA. Under blur, it obtains 54.2% weighted accuracy, clearly outperforming the open-source and tool-augmented baselines in Table 1, including Qwen2.5VL-32B-Instruct(Bai et al., 2025), Qwen3VL-8B-Instruct(Yang et al., 2025a), InternVL2.5-8B(Chen et al., 2025), Qwen2.5VL-7B-Instruct, and Thyme(Zhang et al., 2025c). It also exceeds the SFT+GRPO baseline and Doubao-Seed-1.6, approaches GPT-4.1, and remains the strongest open-source method on clear inputs.

DriveScout shows the smallest clear-to-degraded performance drop on SCOUT. Its Avg. Drop is lower than those of GPT-4.1, Doubao-Seed-1.6, all open-source MLLMs, and the SFT+GRPO baseline. This indicates that DriveScout better preserves small-object understanding across paired clear and motion-degraded observations while the underlying scene and target remain unchanged.

Table 2: Performance comparison with VLM-based driving agents on the perception task of DriveBench under motion blur and zoom blur settings. We report results separately for multi-choice questions (MCQ) and open-ended questions, using GPT-score for evaluation. The best results are highlighted in bold.

Method	Motion Blur		Zoom Blur	
	MCQ	Open-ended	MCQ	Open-ended
Dolphins	9.85	11.34	10.10	10.17
DriveLM	22.98	10.97	20.93	9.03
DriveScout	**43.34**	**34.24**	**46.46**	**32.53**

Table 3: Zero-shot SURDS results on motion-degraded inputs. Only the degradation setting is reported. Best and second-best results are bolded and underlined.

Method	Yaw	Pixel	FB
Doubao-Seed-1.6	4.76	2.17	4.92
GPT-4o	7.62	0.59	8.00
Qwen2.5VL-32B-Instruct	27.84	9.29	**38.86**
Qwen3VL-8B-Instruct	26.22	4.60	34.11
InternVL2.5-8B	<u>30.00</u>	0.30	35.54
Qwen2.5VL-7B-Instruct	22.32	4.15	26.49
Thyme-RL-7B	26.53	<u>24.90</u>	36.46
DriveScout	**34.32**	**69.65**	<u>36.87</u>

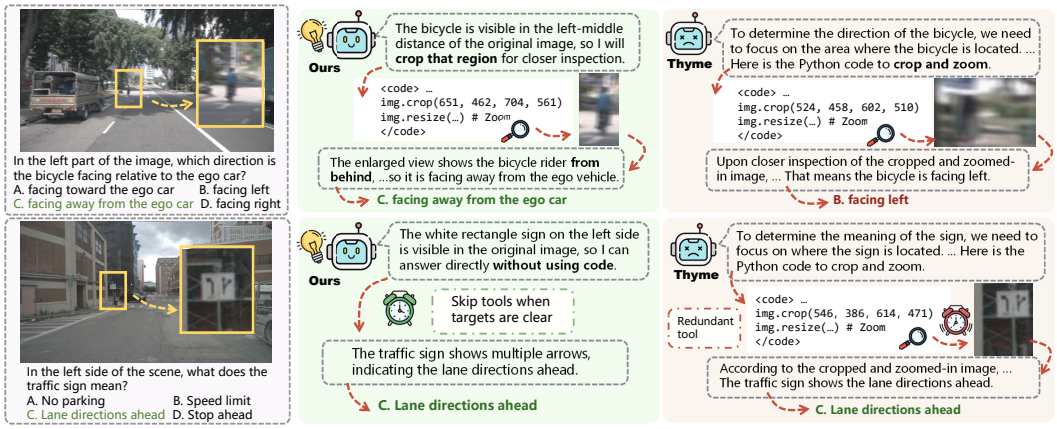


Figure 4: Case study on SCOUT. DriveScout accurately focuses on blurry small targets while avoiding unnecessary visual operations for clear targets, improving both robustness and efficiency.

5.2.2 Performance on SURDS

To verify that our method remains effective beyond SCOUT, we further evaluate it on SURDS(Guo et al., 2024) under the same motion-degradation evaluation protocol. As shown in Table 3, we compare our method with representative open-source and closed-source MLLMs(Hurst et al., 2024) under the zero-shot setting. Our method achieves the best performance on yaw angle classification (Yaw) and pixel-level localization (Pixel), with a particularly strong score of 69.65 on Pixel. This large improvement suggests that standard MLLMs struggle to directly solve blur-degraded tasks requiring precise local localization, while tool-based local visual processing can substantially improve such fine-grained spatial understanding in challenging degraded driving scenes. On front–back understanding (FB), our method also achieves competitive performance.

5.2.3 Performance on DriveBench

We further compare DriveScout with representative VLM-based driving agents on DriveBench perception tasks (Xie et al., 2025) under motion and zoom blur. Under motion blur, it obtains GPT-Scores of 43.34 on MCQ and 34.24 on open-ended questions, outperforming DriveLM (Sima et al., 2024) by 20.36 and 23.27 points, respectively. Under zoom blur, DriveScout also clearly outperforms both DriveLM and Dolphins (Ma et al., 2024). These gains on DriveBench provide complementary evidence beyond SCOUT's paired degradation protocol.

5.3 Diagnostic Analysis

Aggregate benchmark scores alone do not reveal whether the improvement comes from blur-robust localization, more frequent visual operations, or a favorable choice of optimization weights. We therefore examine DriveScout at both the component and behavior levels. The component ablation isolates the contributions of PBC, TGSC, and NAFO, while the hyperparameter analysis measures

how spatial supervision and redundant-focus regularization affect accuracy and crop usage. We additionally inspect representative clear and blurred examples to relate these quantitative trends to the model's focusing decisions.

5.3.1 Ablation Study

Ablation of Key Components. We conduct ablations on SCOUT to evaluate the three components. As shown in Table 4, removing either the Progressive Blur Curriculum (PBC) or Target-Grounded Spatial Credit (TGSC) reduces blurred-setting accuracy, supporting the roles of degradation-aware sampling and spatial supervision. Removing Necessity-Aware Focus Optimization (NAFO) substantially increases crop rate and response length with only a small accuracy change, indicating that counterfactual necessity primarily distinguishes useful local inspection from redundant image operations. Combining the three components yields the best accuracy with a substantially lower crop rate than the unguided base model.

Table 4: Ablation study of key components. PBC: Progressive Blur Curriculum; TGSC: Target-Grounded Spatial Credit; NAFO: Necessity-Aware Focus Optimization. Crop Rate denotes the frequency of crop-based visual focusing, and Tokens denotes the average number of generated tokens per response.

Setting	PBC	TGSC	NAFO	Acc.↑	Crop Rate↓	Tokens↓
Base	–	–	–	48.70	57.30	132.81
w/o PBC	–	✓	✓	52.60	**27.00**	**92.07**
w/o TGSC	✓	–	✓	50.40	37.70	108.13
w/o NAFO	✓	✓	–	<u>53.70</u>	58.90	136.64
DriveScout	✓	✓	✓	**54.20**	<u>28.50</u>	<u>93.51</u>

Ablation of Hyperparameters. As shown in Fig. 5(b), increasing the crop-objective weight initially improves performance, confirming that coordinate-level supervision benefits blur-degraded perception. Larger weights bring no further gains, so crop supervision should remain auxiliary to semantic answering.

As shown in Fig. 5(a), increasing the redundant-focus penalty reduces focusing frequency but may hurt accuracy, revealing an efficiency–accuracy trade-off. A moderate penalty suppresses redundant focusing while preserving performance.

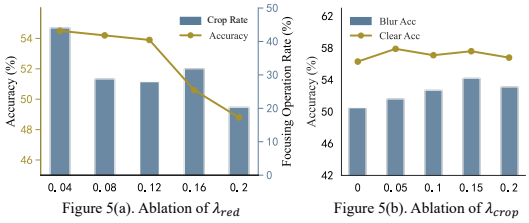


Figure 5(a). Ablation of λ_{red} Figure 5(b). Ablation of λ_{crop}

Figure 5: Hyperparameter sensitivity of (a) the redundant-focus penalty λ_{red} and (b) the crop-objective weight λ_{crop}, showing their effects on accuracy and focusing frequency.

5.3.2 Case Study

We further conduct a case study on SCOUT, as shown in Fig. 4. In the blurred bicycle case, the baseline suffers from *crop drift*, where the focused region shifts away from the target and leads to an incorrect answer. In contrast, DriveScout accurately localizes the bicycle region under motion blur and maintains robust recognition. In the clear traffic-sign case, the baseline still performs unnecessary visual focusing, while DriveScout answers directly without redundant focus operations. These examples show that DriveScout improves both blur-robust localization and focusing efficiency.

6 Conclusion

We study small-scale target understanding for driving MLLMs under parameterized motion degradation and introduce SCOUT, a paired VQA benchmark with matched clear and degraded observations. DriveScout treats visual focusing as selective evidence acquisition and gates coordinate-level supervision by counterfactual necessity, improving performance under the evaluated degradation protocol while reducing redundant operations. Results on SCOUT, SURDS, and DriveBench, together with component ablations, support this design. **Limitations.** SCOUT uses parameterized image-space degradation rather than naturally captured, scene-dependent blur and covers a limited set of target categories. Future work should evaluate real blurred videos, broader scenarios, richer visual operations beyond cropping, and temporally extended interactive driving.

AI USE STATEMENT

GPT-5 was used only to generate reasoning trajectories for the SFT training data. Each generated trajectory was manually checked before use. These trajectories provide training supervision and are not used as evaluation answer labels. Generative AI assistance was also used to migrate the manuscript's LaTeX source to the ICLR 2027 template and to check compilation and formatting; the compilation logs and rendered PDF were inspected during this process.

REPRODUCIBILITY STATEMENT

The appendix describes SCOUT filtering, the paired motion-degradation protocol, and dataset statistics (Appendix A.1), the visual-focusing protocol (Appendix A.2), SFT settings (Appendix A.3), Necessity-Aware Focus Optimization and its reward cases (Appendix A.4), external-benchmark evaluation (Appendix A.5), and RL training settings (Appendix A.6).

REFERENCES

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

Yipene Cedric Francois Bassole, Sungwoo Kim, Jiwoo Jung, and Yunsick Sung. HybridDriveVLA: Vision-language-action model with visual CoT reasoning and ToT evaluation for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 32421–32430, June 2026.

Mario Bijelic, Tobias Gruber, Fahim Mannan, Florian Kraus, Werner Ritter, Klaus Dietmayer, and Felix Heide. Seeing through fog without seeing fog: Deep multimodal sensor fusion in unseen adverse weather. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11682–11692, 2020.

Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11621–11631, 2020.

Xu Cao, Tong Zhou, Yunsheng Ma, Wenqian Ye, Can Cui, Kun Tang, Zhipeng Cao, Kaizhao Liang, Ziran Wang, James M Rehg, et al. Maplm: A real-world large-scale vision-language benchmark for map and traffic scene understanding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 21819–21830, 2024.

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)*, pp. 14093–14100. IEEE, 2024.

Qimao Chen, Fang Li, Yuechen Luo, Zehan Zhang, Haiyang Sun, Fangzhen Li, Bing Wang, Guang Chen, Yang Ji, Jiong Deng, Hongwei Xie, Hangjun Ye, Long Chen, and Yi Zhang. DriveReward: A comprehensive dataset and generative vision-language reward model for autonomous driving, 2026. URL https://arxiv.org/abs/2606.08525.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, Lixin Gu, Xuehui Wang, Qingyun Li, Yiming Ren, Zixuan Chen, Jiapeng Luo, Jiahao Wang, Tan Jiang, Bo Wang, Conghui He, Botian Shi, Xingcheng Zhang, Han Lv, Yi Wang, Wenqi Shao, Pei Chu, Zhongying Tu, Tong He, Zhiyong Wu, Huipeng Deng, Jiaye Ge, Kai Chen, Kaipeng Zhang, Limin Wang, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling, 2025. URL https://arxiv.org/abs/2412.05271.

Gong Cheng, Xiang Yuan, Xiwen Yao, Kebing Yan, Qinghua Zeng, Xingxing Xie, and Junwei Han. Towards large-scale small object detection: Survey and benchmarks. *IEEE transactions on pattern analysis and machine intelligence*, 45(11):13467–13488, 2023.

Tushar Choudhary, Vikrant Dewangan, Shivam Chandhok, Shubham Priyadarshan, Anushka Jain, Arun K Singh, Siddharth Srivastava, Krishna Murthy Jatavallabhula, and K Madhava Krishna. Talk2bev: Language-enhanced bird's-eye view maps for autonomous driving. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 16345–16352. IEEE, 2024.

Charles Corbiere, Simon Roburin, Syrielle Montariol, Antoine Bosselut, and Alexandre Alahi. Retrieval-based interleaved visual chain-of-thought in real-world driving scenarios. *arXiv preprint arXiv:2501.04671*, 2025.

Yinpeng Dong, Caixin Kang, Jinlai Zhang, Zijian Zhu, Yikai Wang, Xiao Yang, Hang Su, Xingxing Wei, and Jun Zhu. Benchmarking robustness of 3d object detection to common corruptions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1022–1032, 2023.

Daocheng Fu, Xin Li, Licheng Wen, Min Dou, Pinlong Cai, Botian Shi, and Yu Qiao. Drive like a human: Rethinking autonomous driving with large language models. In *2024 IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW)*, pp. 910–919. IEEE, 2024.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Xianda Guo, Ruijun Zhang, Yiqun Duan, Yuhang He, Dujun Nie, Wenke Huang, Chenming Zhang, Shuai Liu, Hao Zhao, and Long Chen. Surds: Benchmarking spatial understanding and reasoning in driving scenarios with vision language models. *arXiv preprint arXiv:2411.13112*, 2024.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

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

Ayesha Ishaq, Jean Lahoud, Ketan More, Omkar Thawakar, Ritesh Thawkar, Dinura Dissanayake, Noor Ahsan, Yuhao Li, Fahad Shahbaz Khan, Hisham Cholakkal, et al. Drivelmm-o1: A step-by-step reasoning dataset and large multimodal model for driving scenario understanding. In *2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 20501–20508. IEEE, 2025.

Bo Jiang, Shaoyu Chen, Bencheng Liao, Xingyu Zhang, Wei Yin, Qian Zhang, Chang Huang, Wenyu Liu, and Xinggang Wang. Senna: Bridging large vision-language models and end-to-end autonomous driving. *arXiv preprint arXiv:2410.22313*, 2024.

Jinkyu Kim, Anna Rohrbach, Trevor Darrell, John Canny, and Zeynep Akata. Textual explanations for self-driving vehicles. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 563–578, 2018.

Peizheng Li, Zhenghao Zhang, David Holtz, Hang Yu, Yutong Yang, Yuzhi Lai, Rui Song, Andreas Geiger, and Andreas Zell. SpaceDrive: Infusing spatial awareness into VLM-based autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 40096–40107, June 2026a.

Yiheng Li, Cunxin Fan, Chongjian Ge, Zhihao Zhao, Chenran Li, Chenfeng Xu, Huaxiu Yao, Masayoshi Tomizuka, Bolei Zhou, Chen Tang, et al. Womd-reasoning: A large-scale dataset for interaction reasoning in driving. *arXiv preprint arXiv:2407.04281*, 2024.

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, pp. 6708–6716, 2026b.

Haicheng Liao, Hanlin Kong, Bonan Wang, Chengyue Wang, Kanye Ye Wang, Zhengbing He, Chengzhong Xu, and Zhenning Li. Cot-drive: Efficient motion forecasting for autonomous driving with llms and chain-of-thought prompting. *IEEE Transactions on Artificial Intelligence*, 7(2): 625–641, 2025.

Haicheng Liao, Huanming Shen, Bonan Wang, Yongkang Li, Yihong Tang, Chengyue Wang, Dingyi Zhuang, Kehua Chen, Hai Yang, Chengzhong Xu, and Zhenning Li. Think before you drive: World model-inspired multimodal grounding for autonomous vehicles, 2026. URL https://arxiv.org/abs/2512.03454.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

Yuqi Liu, Tianyuan Qu, Zhisheng Zhong, Bohao Peng, Shu Liu, Bei Yu, and Jiaya Jia. Visionreasoner: Unified visual perception and reasoning via reinforcement learning. *arXiv e-prints*, pp. arXiv–2505, 2025.

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.

Yingzi Ma, Yulong Cao, Jiachen Sun, Marco Pavone, and Chaowei Xiao. Dolphins: Multimodal language model for driving. In *European Conference on Computer Vision*, pp. 403–420. Springer, 2024.

Jiageng Mao, Yuxi Qian, Junjie Ye, Hang Zhao, and Yue Wang. Gpt-driver: Learning to drive with gpt. *arXiv preprint arXiv:2310.01415*, 2023a.

Jiageng Mao, Junjie Ye, Yuxi Qian, Marco Pavone, and Yue Wang. A language agent for autonomous driving. *arXiv preprint arXiv:2311.10813*, 2023b.

Ana-Maria Marcu, Long Chen, Jan Hünermann, Alice Karnsund, Benoit Hanotte, Prajwal Chidananda, Saurabh Nair, Vijay Badrinarayanan, Alex Kendall, Jamie Shotton, et al. Lingoqa: Visual question answering for autonomous driving. In *European Conference on Computer Vision*, pp. 252–269. Springer, 2024.

Ming Nie, Renyuan Peng, Chunwei Wang, Xinyue Cai, Jianhua Han, Hang Xu, and Li Zhang. Reason2drive: Towards interpretable and chain-based reasoning for autonomous driving. In *European Conference on Computer Vision*, pp. 292–308. Springer, 2024.

OpenAI. Thinking with images. https://openai.com/index/thinking-with-images/, 2025. Accessed: 2026-05-06.

Zhenghao Peng, Wenhao Ding, Yurong You, Yuxiao Chen, Wenjie Luo, Thomas Tian, Yulong Cao, Apoorva Sharma, Danfei Xu, Boris Ivanovic, Boyi Li, Yan Wang, and Marco Pavone. Counterfactual VLA: Self-reflective vision-language-action model with adaptive reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4022–4031, June 2026.

Kangan Qian, Sicong Jiang, Yang Zhong, Ziang Luo, Zilin Huang, Tianze Zhu, Kun Jiang, Mengmeng Yang, Zheng Fu, Jinyu Miao, et al. Agentthink: A unified framework for tool-augmented chain-of-thought reasoning in vision-language models for autonomous driving. *arXiv preprint arXiv:2505.15298*, 1(2):3, 2025.

Tianwen Qian, Jingjing Chen, Linhai Zhuo, Yang Jiao, and Yu-Gang Jiang. Nuscenes-qa: A multi-modal visual question answering benchmark for autonomous driving scenario. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 4542–4550, 2024.

Xiaozhen Qiao, Da Zhang, Yubin Guo, Junyu Gao, Zhiyuan Zhao, and Xuelong Li. Semantic-aware temporal adaptation for UAV anti-UAV tracking, 2026. URL https://arxiv.org/abs/2607.26511.

Ishaan Rawal, Shubh Gupta, Yihan Hu, and Wei Zhan. NoRD: A data-efficient vision-language-action model that drives without reasoning, 2026. URL https://arxiv.org/abs/2602.21172.

Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Acdc: The adverse conditions dataset with correspondences for semantic driving scene understanding. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 10765–10775, 2021.

Hao Shao, Yuxuan Hu, Letian Wang, Guanglu Song, Steven L Waslander, Yu Liu, and Hongsheng Li. Lmdrive: Closed-loop end-to-end driving with large language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 15120–15130, 2024.

Chonghao Sima, Katrin Renz, Kashyap Chitta, Li Chen, Hanxue Zhang, Chengen Xie, Jens Beißwenger, Ping Luo, Andreas Geiger, and Hongyang Li. Drivelm: Driving with graph visual question answering. In *European conference on computer vision*, pp. 256–274. Springer, 2024.

Mingzhe Tao, Ruiping Liu, Junwei Zheng, Yufan Chen, Kedi Ying, M. Saquib Sarfraz, Kailun Yang, Jiaming Zhang, and Rainer Stiefelhagen. DriveXQA: Cross-modal visual question answering for adverse driving scene understanding, 2026. URL https://arxiv.org/abs/2603.11380.

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.

Shaoyuan Xie, Lingdong Kong, Yuhao Dong, Chonghao Sima, Wenwei Zhang, Qi Alfred Chen, Ziwei Liu, and Liang Pan. Are vlms ready for autonomous driving? an empirical study from the reliability, data and metric perspectives. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6585–6597, 2025.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025a.

Senqiao Yang, Junyi Li, Xin Lai, Bei Yu, Hengshuang Zhao, and Jiaya Jia. Visionthink: Smart and efficient vision language model via reinforcement learning. *arXiv preprint arXiv:2507.13348*, 2025b.

Seungjun Yu, Junsung Park, Youngsun Lim, and Hyunjung Shim. Robust driving qa through metadata-grounded context and task-specific prompts. *arXiv preprint arXiv:2510.19001*, 2025.

Xuehui Yu, Yuqi Gong, Nan Jiang, Qixiang Ye, and Zhenjun Han. Scale match for tiny person detection. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 1257–1265, 2020.

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.

Lingjun Zhang, Yujian Yuan, Changjie Wu, Xinyuan Chang, Xin Cai, Shuang Zeng, Linzhe Shi, Sijin Wang, Hang Zhang, and Mu Xu. MindDriver: Introducing progressive multimodal reasoning for autonomous driving, 2026. URL https://arxiv.org/abs/2602.21952.

Xintong Zhang, Zhi Gao, Bofei Zhang, Pengxiang Li, Xiaowen Zhang, Yang Liu, Tao Yuan, Yuwei Wu, Yunde Jia, Song-Chun Zhu, et al. Adaptive chain-of-focus reasoning via dynamic visual search and zooming for efficient vlms. *arXiv preprint arXiv:2505.15436*, 2025a.

Xintong Zhang, Zhi Gao, Bofei Zhang, Pengxiang Li, Xiaowen Zhang, Yang Liu, Tao Yuan, Yuwei Wu, Yunde Jia, Song-Chun Zhu, et al. Chain-of-focus: Adaptive visual search and zooming for multimodal reasoning via rl. *arXiv e-prints*, pp. arXiv–2505, 2025b.

Yi-Fan Zhang, Xingyu Lu, Shukang Yin, Chaoyou Fu, Wei Chen, Xiao Hu, Bin Wen, Kaiyu Jiang, Changyi Liu, Tianke Zhang, et al. Thyme: Think beyond images. *arXiv preprint arXiv:2508.11630*, 2025c.

Weicheng Zheng, Xiaofei Mao, Nanfei Ye, Pengxiang Li, Kun Zhan, Xianpeng Lang, and Hang Zhao. Driveagent-r1: Advancing vlm-based autonomous driving with active perception and hybrid thinking. *arXiv preprint arXiv:2507.20879*, 2025a.

Ziwei Zheng, Michael Yang, Jack Hong, Chenxiao Zhao, Guohai Xu, Le Yang, Chao Shen, and Xing Yu. Deepeyes: Incentivizing" thinking with images" via reinforcement learning. *arXiv preprint arXiv:2505.14362*, 2025b.

A SUPPLEMENTARY DETAILS

This appendix supplies the construction, operation, optimization, and evaluation details omitted from the main paper. We organize it around dataset reproducibility, the crop protocol, the role of SFT, Necessity-Aware Focus Optimization, transfer to external benchmarks, and the complete training configuration.

A.1 SCOUT CONSTRUCTION AND STATISTICS

Small-target filtering. We project the nuScenes 3D annotations into each camera view and clip the resulting 2D boxes to the image boundary. Candidate targets are retained using the filters in Table 5. The same-type neighborhood rule removes spatially ambiguous references: the center of the nearest object of the same coarse type must be farther than both an absolute pixel threshold and a box-size-dependent threshold.

Table 5: Candidate-filtering configuration used to construct SCOUT.

Criterion	Retention rule
nuScenes visibility token	≥ 3
Projected area kept after clipping	≥ 0.70
Projected box height	20–80 pixels
Projected box area	≥ 800 pixels^2
Box-to-image area ratio	≤ 0.01
Estimated occlusion ratio	≤ 0.70
Nearest same-type center	$> \max(80 \text{ px}, 1.8\, d_{\text{box}})$
Pedestrian motion threshold	0.25 m/s
Vehicle motion threshold	0.50 m/s

Question-answer construction. Vehicle-heading and pedestrian-motion answers are derived from the corresponding nuScenes annotations, whereas traffic-sign semantics are manually annotated from the source images. All candidate QA instances are subsequently checked against the source image and an enlarged target crop.

Paired motion-degradation protocol. For every retained training example with a matched box, we retain the source image as the clear reference and derive three matched motion-degraded variants. A normalized line kernel is rotated and applied with replicated boundary padding. Table 6 reports both the sampling parameters and the realized odd kernel sizes in the stored training set. The driving-direction prior assigns probability 0.85 to near-horizontal motion, with the remaining probability distributed over diagonal, vertical, and unrestricted directions. With probability 0.08, we also add weak Gaussian sensor noise with standard deviation sampled uniformly from $[0.8, 2.0]$.

Table 6: SCOUT motion-blur configuration. The random seed is 42, the angular jitter is $\pm 6^\circ$, and all three levels use the driving-direction prior.

Variant	Scale	Realized kernel sizes	Number of samples
Clear	–	–	3,718
Light blur (ℓ_1)	0.35	$\{5, 7, 9\}$	3,718
Medium blur (ℓ_2)	0.45	$\{7, 9, 11\}$	3,718
Strong blur (ℓ_3)	0.55	$\{9, 11, 13\}$	3,718

Human verification. Annotators inspect the original image together with an enlarged target crop. They reject an example if the referenced object is not visible, the annotated answer cannot be verified visually, another plausible object makes the reference ambiguous, or the crop lacks sufficient target context. This step checks semantic answerability rather than merely rechecking the automatically projected box.

Split and task distribution. The training and test sets are scene-disjoint according to the nuScenes scene identifiers. The stored release contains 3,769 training and 1,000 test QA pairs. Of the training pairs, 3,718 have a matched box and are expanded into the four visual variants above, yielding 14,872 RL records. The four QA templates cover three target families: vehicle understanding, pedestrian understanding, and traffic sign understanding; the first two use nuScenes-derived answers, while sign semantics are manually annotated.

A.2 Visual-Focusing and Crop Protocol

Allowed operations and operation budget. DriveScout permits crop-based visual focusing during both training and inference. A crop may be resized before being returned to the MLLM, but it does not alter the underlying answer label. We use at most two successful crop operations per sample in both phases; after the budget is exhausted, the model must produce its final answer from the accumulated visual evidence.

No-crop counterfactual. The direct-view check used for counterfactual evidence necessity estimation is invoked only for a correct, focus-assisted rollout. It reuses the original image and question but removes crop and sandbox outputs, then appends the following instruction:

```
Now answer the question directly WITHOUT any <code> or
<sandbox_output>.  Do NOT write code.  Do NOT mention tools.
Output strictly:  <think>...</think> <answer>...</answer>.
```

The counterfactual uses greedy decoding (temperature 0 and top-$p = 1$). Crop and sandbox tags are also registered as stop strings, so a generated crop request cannot be executed in this branch. Thus, a_i^{dv} measures whether the same policy can solve the sample using the full image alone.

A.3 SFT Data, Settings, and Diagnostic Effect

The SFT mixture contains 6,313 reasoning trajectories: 2,599 include an image operation and 3,714 answer directly. This mixture is deliberate: crop trajectories teach the syntax and spatial pattern of local inspection, while direct trajectories prevent the model from treating a tool call as mandatory. Table 7 gives the complete SFT configuration.

Table 7: SFT configuration.

Item	Setting
Initialization	Thyme-RL-7B
Optimization scope	Full language model; vision encoder frozen
Precision / attention	bfloat16 / FlashAttention
Epochs	1
Learning rate / warmup	1×10^{-6} / 0.05
Batch size	1 per GPU
Gradient accumulation	4
Maximum sequence length	10,240 tokens
Distributed optimizer	DeepSpeed ZeRO-2
Hardware	8 NVIDIA H20 GPUs (96 GB each)

As an SFT-only diagnostic on the 1,000-example clear test set, the routed SFT checkpoint obtains 54.7% accuracy, invokes crop-based processing on 55.3% of samples, and generates 132.81 tokens per sample on average. The subsequent RL stage is therefore responsible not only for further accuracy gains but also for calibrating when this learned operation should be used. Table 4 in the main paper reports the resulting accuracy–efficiency trade-off.

A.4 Details of Necessity-Aware Focus Optimization

Equation 3 distinguishes four mutually exclusive outcomes, shown in Table 8. Correct direct answers receive a small positive reward. A correct focus-assisted answer receives a positive reward only if the forced direct-view answer is wrong, establishing that local evidence was necessary. If the

direct-view answer is also correct, the focus operation is redundant and is penalized. Finally, an incorrect focus-assisted answer is penalized regardless of the counterfactual.

Table 8: Case-wise interpretation of Eq. 3.

f_i	a_i^{main}	a_i^{dv}	Focus reward
0	1	–	$+\lambda_{\text{dir}}$ (correct direct answer)
0	0	–	0 from the focus term
1	1	0	$+\lambda_{\text{nec}}$ (necessary focus)
1	1	1	$-\lambda_{\text{red}}$ (redundant focus)
1	0	–	$-\lambda_{\text{fail}}$ (failed focus)

In the implemented configuration, $\lambda_{\text{nec}} = 0.20$, $\lambda_{\text{red}} = 0.06$, $\lambda_{\text{fail}} = 0.05$, and $\lambda_{\text{dir}} = 0.08$. Direct-view checking is applied to all eligible correct focus-assisted rollouts, with at most eight checks per batch. When a focus operation is judged redundant, its crop-coordinate advantage and token mask are set to zero, so the model is not trained to imitate the coordinates of an unnecessary crop.

A.5 External-Benchmark Blur Protocol and Zero-Shot Results

SURDS adaptive blur. We apply the same normalized motion-line kernel and driving-direction prior to SURDS. The processing plan retains 20% of samples as clear. Of the remaining 80%, 30% use scale 0.35, 50% use scale 0.45, and 20% use scale 0.55. Equivalently, the full evaluation mixture contains 20% clear, 24% light, 40% medium, and 16% strong blur. The angle jitter, horizontal bias, sensor noise probability, and kernel construction match SCOUT.

DriveBench. We use the official motion-blur and zoom-blur corruptions and evaluate only the perception subset. All entries are zero-shot: no SURDS or DriveBench sample is used in SFT or RL.

A.6 Complete RL Training Configuration

We initialize GRPO from the one-epoch SFT checkpoint and train all language-model parameters for five epochs. The vision encoder remains fixed. Table 9 lists all optimization and generation parameters, and Table 10 gives the stage-wise sampling distribution.

Table 9: GRPO training and rollout configuration.

Item	Setting
Training data	3,718 bases $\times$ 4 visual variants
Epochs / curriculum stages	5 / $(2, 2, 1)$ epochs
Responses per prompt	8
Batch size / gradient accumulation	1 per GPU / 5
Learning rate	5×10^{-7}
Scheduler / warmup	Cosine, minimum LR ratio 0.1 / 0.03
KL coefficient β	0.01
Maximum sequence / completion	8,192 / 768 tokens
Sampling	$T = 1.0$, top-$p = 0.9$, top-$k = 50$
Repetition penalty	1.05
Maximum images per prompt	6
Crop-objective weight λ_{crop}	0.15
Training precision	bfloat16
Distributed optimizer	DeepSpeed ZeRO-3 with offloading
Rollout engine	colocated vLLM, tensor parallel size 1
vLLM memory utilization	0.55
Attention implementation	FlashAttention
Hardware	8 NVIDIA H20 GPUs (96 GB each)

Table 10: Progressive Blur Curriculum probabilities.

Stage	Epochs	Clear	ℓ_1	ℓ_2	ℓ_3
1	2	0.60	0.25	0.10	0.05
2	2	0.35	0.25	0.25	0.15
3	1	0.20	0.20	0.30	0.30