

# Enhancing VLMs for Autonomous Driving through Task-Specific Prompting and Spatial Reasoning

Aodi Wu<sup>1,2</sup>, Xubo Luo<sup>1,2</sup>

<sup>1</sup>University of Chinese Academy of Sciences

<sup>2</sup>Technology and Engineering Center for Space Utilization, Chinese Academy of Sciences  
Beijing, China

wuaodi20@mailsucas.ac.cn, luoxubo23@mailsucas.ac.cn

## Abstract

*This technical report presents our solution for the RoboSense Challenge at IROS 2025, which evaluates Vision-Language Models (VLMs) on autonomous driving scene understanding across perception, prediction, planning, and corruption detection tasks. We propose a systematic framework built on four core components. First, a Mixture-of-Prompts router classifies questions and dispatches them to task-specific expert prompts, eliminating interference across diverse question types. Second, task-specific prompts embed explicit coordinate systems, spatial reasoning rules, role-playing, Chain-of-Thought/Tree-of-Thought reasoning, and few-shot examples tailored to each task. Third, a visual assembly module composes multi-view images with object crops, magenta markers, and adaptive historical frames based on question requirements. Fourth, we configure model inference parameters (temperature, top-p, message roles) per task to optimize output quality. Implemented on Qwen2.5-VL-72B, our approach achieves 70.87% average accuracy on Phase-1 (clean data) and 72.85% on Phase-2 (corrupted data), demonstrating that structured prompting and spatial grounding substantially enhance VLM performance on safety-critical autonomous driving tasks. Code and prompt are available at <https://github.com/wuaodi/UCAS-CSU-phase2>.*

## 1. Introduction

Autonomous driving systems require accurate scene understanding to make safe decisions. With the advancement of Vision-Language Models (VLMs) [1–7], there is growing interest in applying these models to autonomous driving tasks [8–10]. The RoboSense Challenge at IROS 2025 provides a comprehensive benchmark spanning perception, prediction, planning, and sensor and data corruption detection [11, 12].

Despite their strong general capabilities, applying VLMs

to autonomous driving faces three critical challenges [8, 13]. First, *spatial reasoning in multi-view scenarios* is particularly difficult – VLMs often confuse left/right directions, misinterpret objects from back-view cameras as being ahead of the ego vehicle, and struggle with coordinate-based object references [9, 14]. Second, *prompt interference across diverse task types* degrades performance when a single universal prompt attempts to cover perception, prediction, planning, and corruption detection simultaneously [8, 15]; optimizing the prompt for one task often hurts others. Third, *temporal context integration* requires careful design – naively adding historical frames can introduce noise and distraction rather than helpful motion cues, especially when the temporal evidence is irrelevant to the specific question.

We address these challenges through three key design principles. First, we eliminate prompt interference by routing each question to a specialized expert prompt rather than using a single universal instruction. Second, we enhance spatial reasoning by explicitly defining multi-view coordinate systems and domain-specific constraints (e.g., BACK-camera objects are always behind the ego vehicle) that serve as grounding anchors. Third, we adapt temporal context per question type, selecting appropriate historical frame modes to provide relevant motion cues while avoiding distraction. These principles are realized through a four-component pipeline: a router, task-specific prompts with structured reasoning (Chain-of-Thought/Tree-of-Thought), adaptive visual assembly, and per-task inference parameters. Implemented on Qwen2.5-VL-72B, our system achieves 70.87% on Phase-1 and 72.85% on Phase-2.

## 2. Related Work

### 2.1. VLMs for Autonomous Driving

Recent technical reports advance the general capability, reasoning and efficiency of open-source VLMs, such as Qwen2.5-VL [3], GLM-4.5V/4.1V-Thinking [16], and InternVL3.5 [17]. These models provide strong backbones for

driving-oriented perception and reasoning [9, 18, 19].

From the evaluation perspective, DriveBench [8] surfaces reliability issues highly relevant to safety-critical autonomy: weak visual grounding, sensitivity to corruptions, and multi-modal reasoning gaps [15, 20–23]. Ego3D-Bench [24] further shows that multi-view, ego-centric spatial reasoning remains challenging for current VLMs and proposes Ego3D-VLM to inject structured spatial priors. Our method is complementary: instead of finetuning models, we target reliability by prompt-level design – explicit spatial grounding, adaptive temporal evidence, and task-specific routing – so that strong VLMs can be better aligned to autonomous-driving questions.

## 2.2. Prompt Engineering

Prompt engineering offers controllable ways to elicit reasoning and stabilize outputs. Role-playing provides a principled lens to steer behaviour [25], and RoleLLM [26] shows that explicit role profiles and role-conditioned tuning can enhance style control. Chain-of-Thought prompting elicits step-by-step reasoning [27]; Tree-of-Thought extends this to explore multiple paths with self-evaluation [28]; self-reflection further improves problem solving via post-hoc critique [29]. In-context learning [30, 31] shapes format and priors through compact exemplars, while visual prompting surveys techniques to ground attention on images [32]. Beyond a single instruction, Mixture-of-Prompts [33, 34] motivates using multiple expert prompts.

Guided by these insights, our system instantiates: (i) a router that dispatches queries to *task-specific* prompts; (ii) an expert prompt design combining a coordinate-system block, reasoning guidance (CoT/ToT with brief self-check), and few-shot examples; and (iii) visual prompting with markers and crops plus adaptive history. This design leverages strong foundation VLMs while directly addressing reliability and spatial-temporal grounding concerns highlighted by prior studies.

## 3. Methodology

Our approach leverages systematic prompt engineering to enhance VLM performance on autonomous driving scene understanding. As illustrated in Figure 1, the pipeline consists of four core components: a *Router* that classifies each test query and dispatches it to the appropriate task-specific expert; *Task-Specific Prompts* that embed coordinate systems, spatial reasoning rules, role-playing, Chain-of-Thought/Tree-of-Thought reasoning, and few-shot examples tailored to each question type; a *Visual Assembly* module that composes multi-view images, object crops with optional magenta markers, and adaptive historical frames; and finally the *VLM* (Qwen2.5-VL) with task-dependent inference parameters that generates the answer.

A key insight underlying all components is that explicit coordinate system definitions and spatial reasoning constraints are essential to ground VLMs in multi-view autonomous driving scenarios, where implicit scene understanding often fails. The following subsections detail each component in turn.

### 3.1. Router

A critical design choice in our system is to route questions to task-specific prompts [33] rather than using a single universal prompt. Early experiments revealed that when a unified prompt attempted to cover all question types with diverse instructions and examples, performance gains on some categories were often accompanied by drops on others. We attribute this to *prompt confusion*: overly long or heterogeneous prompts can cause the model to struggle with conflicting guidance, leading to unstable behavior.

To mitigate this, we implement a lightweight router that classifies each query into one of seven supported types (Perception-MCQs, Perception-VQA-Object, Perception-VQA-Scene, Prediction-MCQs, Planning-VQA-Object, Planning-VQA-Scene, and Corruption-MCQs) and dispatches it to the corresponding specialized prompt. Given that the challenge questions follow fixed templates, we employ a rule-based approach leveraging the question category field provided in the dataset JSON and regular expressions to detect object references (e.g., `<c2, CAM_VIEW, x, y>`). This simple strategy proves effective for the structured benchmark setting; for more open-ended or diverse question distributions, one could replace the router with a small LLM-based classifier.

Beyond prompt selection, the router also determines the visual assembly configuration (e.g., which history mode to use) and inference parameters, ensuring each question type receives tailored processing.

### 3.2. Task-Specific Prompts

Each question type is paired with a specialized prompt template. The templates share a compact and interpretable structure that are selected by the router at run time:

**Coordinate system and spatial rules.** We explicitly define the six-camera layout and the  $1600 \times 900$  pixel frame with origin at the top-left. Simple yet effective heuristics guide spatial grounding (e.g.,  $x < 800$  for left-half;  $y < 450$  for upper region). Domain constraints further reduce errors: views from the BACK cameras are always behind the ego; motion is interpreted by orientation and road geometry; lane context hints moving vs. parked objects; and common driving conventions apply (e.g., double yellow lines as dividers, explicit evidence before declaring “stopped”).

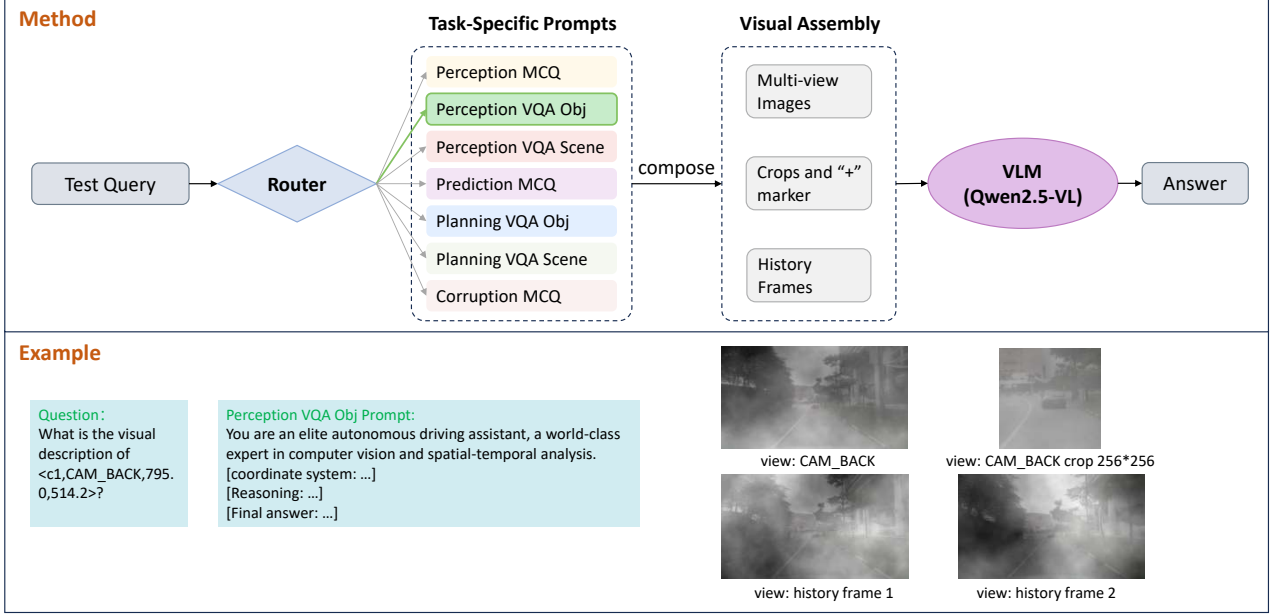


Figure 1. Method overview. A router classifies the test query and selects one task-specific expert (prompt) from multiple candidates. Each expert (shown in different colors) handles a specific question type with specialized instructions and examples. The activated expert’s prompt is then combined with multi-view images, magenta “+” markers, region crops, and adaptive historical frames before invoking the VLM to produce the answer.

**Role-playing.** Each prompt assigns the model a task-specific expert role [25, 26] to guide its reasoning style and establish the appropriate analytical mindset. For instance, perception tasks cast the model as an “elite autonomous driving assistant and world-class expert in computer vision and spatio-temporal analysis”; planning tasks assign it the role of a “world-class autonomous driving risk analyst and strategic planning expert”; while prediction tasks position it as a specialist in “motion forecasting and interaction prediction.” This role-playing mechanism leverages the model’s capacity for in-context persona adaptation, helping to elicit domain-relevant reasoning patterns and maintain consistency in output tone.

**Chain-of-Thought reasoning.** Instead of directly producing an answer, prompts instruct the model to reason step-by-step [27]. The reasoning flow is structured as a sequence: (1) identify static properties from the current frame (object type, color, spatial position); (2) verify dynamic state by comparing position across historical frames to distinguish moving vs. parked objects; (3) synthesize spatial and temporal cues to form a hypothesis; and (4) perform a brief self-check to ensure consistency with domain rules (e.g., BACK-camera constraint, lane-context prior). For VQA tasks the reasoning is exposed in the output, providing interpretability; for MCQs it remains implicit but still guides the internal decision process.

**Tree-of-Thought exploration.** When the query admits multiple plausible outcomes (e.g., predicting whether an object will enter the ego’s path), we apply a lightweight Tree-of-Thought strategy [28]. The prompt explicitly asks the model to consider both branches and then determine which hypothesis is best supported by the visual evidence and motion trends. This dual-path exploration reduces premature commitment and improves robustness on ambiguous cases, albeit at modest additional computational cost.

**Few-shot in-context examples.** Each prompt includes 2–5 carefully curated examples [30] that demonstrate the expected reasoning structure, evidence synthesis, and answer format. These examples show how to reference specific views (e.g., “CAM\_BACK”), cite temporal changes (e.g., “comparing T=0 and T-1 reveals...”), and integrate domain constraints. By anchoring the model’s behavior to concrete demonstrations, we achieve more stable outputs and reduce formatting errors, particularly for complex multi-step questions.

**Prompt example.** Figure 2 shows a condensed Perception-VQA-Object prompt template with key techniques highlighted in color.

**[Role-playing]** You are an elite autonomous driving assistant, a world-class expert in computer vision and spatio-temporal analysis.

**[Coordinate System & Spatial Rules]**

Six cameras: FRONT, FRONT.LEFT/RIGHT, BACK, BACK.LEFT/RIGHT. Image frame: 1600×900, origin top-left.

Spatial heuristics:  $x < 800$  left-half,  $x > 800$  right-half;  $y < 450$  upper,  $y > 450$  lower.

Critical constraint: Any object in BACK cameras is ALWAYS behind the ego vehicle.

Motion rule: Determine parked vs. moving by comparing position across frames (T=0, T-1, T-2).

**[Chain-of-Thought Structure]**

Your reasoning MUST follow: (1) Identify static properties from current frame & crop; (2) Verify dynamic state via temporal comparison; (3) Synthesize evidence; (4) Self-check consistency with spatial rules.

**[Few-shot Example]**

Question: What is `<c2, CAM.BACK.RIGHT, 1324, 552>?`

Reasoning: Object in CAM.BACK.RIGHT  $\Rightarrow$  behind & right of ego. Crop: dark sedan in parking area. Temporal check: position vs. fence unchanged across T=0/T-1/T-2  $\Rightarrow$  parked. Self-check: BACK constraint satisfied.

Final Answer: The object is a dark-colored sedan confirmed parked behind the ego vehicle.

Figure 2. Prompt template excerpt with techniques color-coded: **role-playing**, **coordinate grounding**, **CoT structure**, **in-context example**.

### 3.3. Visual Assembly

Once the router selects a task-specific prompt, the system composes the visual inputs to accompany the textual query. This assembly adapts to the question type and the presence of object references, ensuring each query receives contextually appropriate evidence.

**Multi-view images.** All six camera views are transmitted in a canonical order (FRONT, FRONT.RIGHT, FRONT.LEFT, BACK, BACK.RIGHT, BACK.LEFT) to maintain consistent spatial correspondence across queries. For object-centric questions, only the views containing the referenced objects are sent to reduce noise; for scene-level questions, the full six-view set provides comprehensive spatial context.

**Object-centric visual grounding.** When a query references a specific object via coordinates (e.g., `<c2, CAM.BACK.RIGHT, x, y>`), we augment the full view with two additional cues [32]. First, a magenta “+” marker is optionally overlaid at the object’s location in the

full image to facilitate visual alignment. Second, a  $256 \times 256$  crop centered at the coordinates is appended to provide fine-grained detail that may be lost in the downscaled full view. This dual representation—context plus detail—mirrors human strategies of alternating between scene overview and focused inspection.

**Adaptive temporal context.** Historical frames supply motion cues essential for prediction and planning. We support three modes selected by question type: *grid* mode assembles all six views from each historical timestep into a labeled mosaic, ideal for scene-level reasoning; *front* mode sends only the CAM.FRONT view, concentrating on forward-driving context; and *referenced* mode restricts history to the views containing the query’s referenced object, minimizing irrelevant content. Typically, two historical frames (T-1 and T-2) precede the current frame, though corruption detection disables history to focus on single-frame artifact identification.

**Corruption detection with reference images.** For Corruption-MCQs, a reference image (Figure 3) is prepended to the input. This reference image provides a visual baseline that helps the model distinguish normal appearance from degraded conditions such as fog, blur, or saturation.

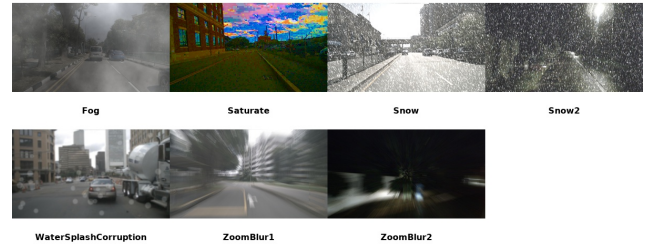


Figure 3. The reference image used for corruption detection.

### 3.4. Model Selection and Inference Parameters

We adopt Qwen2.5-VL-72B-Instruct as our primary backbone due to its strong performance on vision-language reasoning benchmarks and its availability as an open-source model. Beyond prompt engineering, we further tailor the model’s behavior through task-specific inference configurations.

For tasks requiring deterministic and precise outputs—such as multiple-choice questions and object-level descriptions—we employ conservative decoding with low temperature (0.2) and low sampling threshold (top- $p=0.2$ ). This setting suppresses stochastic variation and yields stable, focused responses. Conversely, for scene-level descriptions where richness and contextual nuance are valued, we relax



these constraints (temperature 1.5, top- $p=0.9$ ) to encourage broader exploration of the output space.

An additional refinement involves message role assignment. For most question types, the task-specific prompt is delivered as a *system* message. However, we empirically found that Planning-VQA-Object achieves better performance when the prompt is sent as a *user* message instead, likely due to improved output format compliance.

To support rapid experimentation, we implement a parallelized inference pipeline with fault-tolerant checkpointing. Samples are processed concurrently across multiple workers, with intermediate results saved atomically to prevent data loss. Selective update modes allow re-processing of specific question types or index ranges without re-running the entire dataset, accelerating the iteration cycle for prompt refinement.

## 4. Experiments

### 4.1. Task & Dataset

We use the official data provided by the *RoboSense Challenge 2025* [11] held at IROS 2025. This competition builds upon the legacy of the *RoboDepth Challenge 2023* [35, 36] at ICRA 2023 and the *RoboDrive Challenge 2024* [12, 20] at ICRA 2024, continuing the collective effort to advance robust and scalable robot perception. Each track in this competition is grounded on an established benchmark designed for evaluating real-world robustness and generalization [8, 37–42]. Specifically, this task is built upon the **DriveBench** dataset [8] in **Track 1**, which evaluates vision-language models in autonomous driving through perception, prediction, and planning questions under both clean and corrupted visual conditions.

We evaluate on two competition phases with distinct datasets. *Phase-1* uses clean images and covers six question types: Perception-MCQs, Perception-VQA-Object, Perception-VQA-Scene, Prediction-MCQs, Planning-VQA-Object, and Planning-VQA-Scene. *Phase-2* uses corrupted images and adds a seventh type, Corruption-MCQs.

### 4.2. Implementation Details.

Unless stated otherwise, we use Qwen2.5-VL-72B-Instruct with task-specific decoding parameters (Section 3). Inference runs through a remote API with parallel workers. All reported results are averaged scores across test samples in each question category.

### 4.3. Ablation Study

We present ablation studies for both competition phases to demonstrate the incremental benefit of each design component. In Phase-1, we explore the design space on clean images, systematically building up the core framework from a naive baseline to a robust prompting system. Starting

from a simple six-view concatenation, we progressively introduce crops with coordinate grounding, domain-specific rules, Chain-of-Thought reasoning, and finally the Mixture-of-Prompts router. In Phase-2, we inherit the MoP backbone established in Phase-1. The Phase-2 baseline achieves an average of 58.96—further refine each task-specific prompt, add adaptive temporal context, incorporate corruption-detection mechanisms, and tune model parameters. Tables 1 and 2 show the progressive improvements across both phases, with colored entries highlighting which metrics each technique improves.

#### 4.3.1 Phase-1: Building the Prompting Framework

Key observations from Phase-1 ablation: (1) *Visual grounding is critical*—adding crops and explicit coordinate system definitions yields the largest initial gain (+7.05 average), with particularly strong improvements on Planning-VQA-Object (+10.42) and Plan-Scene (+10.73), confirming that spatial understanding is a primary bottleneck for autonomous driving reasoning tasks. (2) *Domain constraints prevent systematic errors*—while domain rules/constraints provide modest average gains (+0.95), they are essential for eliminating category-level mistakes such as describing BACK-camera objects as “ahead” or predicting unrealistic lane changes. This technique improves Perception-MCQs from 64.15 to 75.47. (3) *CoT unlocks reasoning capability*—switching from result-first to reason-first delivers another substantial jump (+7.83), achieving perfect scores on Perception-MCQs (100.00) and lifting Planning-VQA-Object to 79.79. This validates the importance of explicit reasoning steps. (4) *MoP eliminates prompt interference*—the final router-based Mixture-of-Prompts strategy addresses prompt confusion across diverse question types, notably boosting Perception-VQA-Object (+7.04) and Perception-VQA-Scene (+13.36), culminating in an overall average of 70.87.

#### 4.3.2 Phase-2: Task-Specific Refinements on Corrupted Data

Notable findings from Phase-2 ablation: (1) *Dynamic history mode provides temporal context*—adaptively selecting historical frames based on question type (grid for general awareness, front-only for motion prediction, and referenced for specific inquiries) improves multiple categories, with the average rising from 58.96 to 60.32 (+1.36). This indicates that temporal information helps, but must be task-aligned to avoid distraction. (2) *Task-specific prompt refinement is highly effective*—individually optimizing the Perception-VQA-Object prompt (adding stricter directional constraints and implicit reasoning) brings that category from 37.95 to 50.30, a +12.35 gain, demonstrating the value of tailored instructions. (3) *Corruption detection benefits from reference*

Table 1. Phase-1 ablation study (clean data; 6 types). Each row adds one technique to the previous configuration. Colored entries show improvements brought by that technique.

Configuration	P-MCQ	P-Obj	P-Scene	Pred-MCQ	Plan-Obj	Plan-Scene	Avg	$\Delta$
Baseline (six views + Q)	56.60	38.59	69.92	59.77	58.44	49.83	53.79	–
+ Crops & coordinate system	64.15	46.89	63.12	64.37	68.86	60.56	60.84	+7.05
+ Domain rules/constraints	75.47	43.32	63.12	66.86	66.83	62.40	61.79	+0.95
+ CoT reasoning (reason-first)	100.00	57.43	62.81	72.99	79.79	64.83	69.62	+7.83
+ Mixture-of-Prompts (MoP)	100.00	64.47	76.17	72.99	75.21	64.64	70.87	+1.25

Table 2. Phase-2 ablation study (corruption data; 7 types). Each row adds one technique to the previous configuration. Colored entries show improvements brought by that technique.

Configuration	P-MCQ	P-Obj	P-Scene	Pred-MCQ	Plan-Scene	Plan-Obj	Corr-MCQ	Avg	$\Delta$
Baseline (MoP backbone)	93.88	37.95	62.20	67.48	61.93	48.00	97.12	58.96	–
+ Dynamic history mode	98.98	39.39	63.60	67.83	64.40	50.06	96.15	60.32	+1.36
+ Perception-Obj refinement	98.98	50.30	63.90	67.83	64.50	50.37	96.15	62.31	+1.99
+ Corruption reference image	98.98	50.29	63.71	67.83	64.32	50.12	100.00	62.38	+0.07
+ Tree-of-Thought (ToT)	100.00	50.23	63.80	95.47	63.00	49.37	100.00	71.64	+9.26
+ Task-specific decoding params	100.00	49.92	64.24	95.47	64.12	49.90	100.00	72.01	+0.37
+ Planning-Obj as user role	100.00	49.97	64.51	95.47	64.20	53.10	100.00	72.85	+0.84

images—adding the clean reference image for Corruption-MCQs achieves perfect scores (100.00), validating our visual assembly strategy for corruption-aware tasks. (4) *ToT yields the largest improvement*—incorporating Tree-of-Thought reasoning (exploring multiple hypotheses before conclusion) produces a dramatic +9.26 gain, especially on Prediction-MCQs (67.48  $\rightarrow$  95.47), where uncertainty is high. (5) *Parameter tuning and role assignment refine outputs*—task-specific temperature/top- $p$  values and switching Planning-VQA-Object prompts to user-role further improve format compliance, bringing the final average to 72.85.

#### 4.4. Model Comparisons

We compare Qwen2.5-VL-72B with other state-of-the-art VLMs on both competition phases. Table 3 reports results under identical prompting conditions.

In Phase-1, while GLM-4.5V excels on Perception-VQA-Object (51.48 vs. 46.89), Qwen2.5-VL shows stronger overall consistency and significantly outperforms on Planning-VQA-Object (68.86 vs. 37.34). In Phase-2, Qwen2.5-VL-72B continues to outperform InternVL-3.5 on average (58.96 vs. 54.31), with particularly strong advantages on Perception-VQA-Object and Corruption-MCQs. We therefore adopt Qwen2.5-VL as our primary backbone.

## 5. Conclusion

This technical report presents a systematic approach to autonomous driving scene understanding using Vision-Language Models. We address the challenges of the Ro-

boSense Challenge through task-specific prompt engineering, spatial reasoning enhancement, and adaptive visual assembly.

### 5.1. Key Contributions and Insights

This work presents three main contributions for autonomous driving scene understanding with VLMs. First, we introduce a *Mixture-of-Prompts* router that classifies questions and dispatches task-specific prompts to eliminate interference across diverse question types. Second, we establish *coordinate grounding* through explicit camera layouts, spatial rules (e.g., back-camera constraints), and visual markers that significantly improve multi-view reasoning. Third, we develop *structured reasoning strategies* including Chain-of-Thought for step-by-step analysis and Tree-of-Thought for exploring alternative hypotheses in uncertain scenarios. Our experiments reveal three critical insights: spatial understanding remains a bottleneck for VLMs but can be substantially improved through explicit coordinate systems and domain constraints; reasoning-first approaches consistently outperform direct-answer strategies; and task-specific prompt optimization is more effective than generic instructions when combined with appropriate visual and temporal context.

### 5.2. Limitations and Future Work

Our approach has three main limitations. First, autonomous driving is inherently a temporal task, yet our current implementation uses only two historical frames; longer video sequences may be necessary to capture motion trends and de-

Table 3. Model comparison across both phases under identical prompts and inference settings; best per column in **bold**.

Phase	Model	P-MCQ	P-Obj	P-Scene	Pred-MCQ	Plan-Scene	Plan-Obj	Corr-MCQ	Avg
Phase-1	Qwen2.5-VL-72B	<b>64.15</b>	46.89	<b>63.12</b>	<b>64.37</b>	<b>60.56</b>	<b>68.86</b>	–	<b>60.84</b>
	GLM-4.5V	37.74	<b>51.48</b>	53.98	57.85	52.15	37.34	–	51.93
Phase-2	Qwen2.5-VL-72B	93.88	<b>37.95</b>	<b>62.20</b>	<b>67.48</b>	<b>61.93</b>	<b>48.00</b>	<b>97.12</b>	<b>58.96</b>
	InternVL-3.5	<b>94.90</b>	34.26	60.41	61.79	55.43	43.67	91.35	54.31

rive better driving strategies. Second, Chain-of-Thought and Tree-of-Thought reasoning significantly increase inference costs, limiting practical deployment. Third, spatial understanding challenges – particularly in multi-view geometry – may require solutions at the model training or fine-tuning stage rather than relying solely on prompt engineering.

Future work should explore long-term temporal modeling with extended video contexts, efficient reasoning mechanisms to reduce computational overhead, and VLM fine-tuning with autonomous driving data to fundamentally enhance spatial reasoning capabilities. Our results demonstrate that structured prompting substantially improves VLM performance on autonomous driving tasks, providing a foundation for these future directions.

## References

- [1] OpenAI. Gpt-4v(ision) system card. [https://cdn.openai.com/papers/GPTV\\_System\\_Card.pdf](https://cdn.openai.com/papers/GPTV_System_Card.pdf), 2023. Accessed: September 25, 2023.
- [2] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *NeurIPS*, 2023.
- [3] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.
- [4] Runnan Chen et al. Towards label-free scene understanding by vision foundation models. In *Advances in Neural Information Processing Systems*, volume 36, pages 75896–75910, 2023.
- [5] Lingdong Kong et al. LargeAD: Large-scale cross-sensor data pretraining for autonomous driving. *arXiv preprint arXiv:2501.04005*, 2025.
- [6] Sicheng Feng, Song Wang, Shuyi Ouyang, et al. Can MLLMs guide me home? a benchmark study on fine-grained visual reasoning from transit maps. *arXiv preprint arXiv:2505.18675*, 2025.
- [7] Jingyi Xu et al. Visual foundation models boost cross-modal unsupervised domain adaptation for 3d semantic segmentation. *IEEE Transactions on Intelligent Transportation Systems*, 26(11):20287–20301, 2025.
- [8] Shaoyuan Xie et al. Are VLMs ready for autonomous driving? an empirical study from the reliability, data, and metric perspectives. In *IEEE/CVF International Conference on Computer Vision*, pages 6585–6597, 2025.
- [9] Chonghao Sima, Katrin Renz, Kashyap Chitta, Li Chen, Hanxue Zhang, Chengen Xie, Jens Beßwenger, Ping Luo, Andreas Geiger, and Hongyang Li. DriveLM: Driving with graph visual question answering. In *European conference on computer vision*, pages 256–274. Springer, 2024.
- [10] Lingdong Kong, Xiang Xu, Jiawei Ren, et al. Multi-modal data-efficient 3D scene understanding for autonomous driving. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(5):3748–3765, 2025.
- [11] Lingdong Kong, Shaoyuan Xie, Zeyang Gong, Ye Li, Meng Chu, Ao Liang, Yuhao Dong, Tianshuai Hu, Ronghe Qiu, Rong Li, Hanjiang Hu, Dongyue Lu, Wei Yin, Wenhao Ding, Linfeng Li, Hang Song, Wenwei Zhang, Yuexin Ma, Junwei Liang, Zhedong Zheng, Lai Xing Ng, Benoit R. Cottareau, Wei Tsang Ooi, Ziwei Liu, Zhanpeng Zhang, Weichao Qiu, Wei Zhang, Ji Ao, Jiangpeng Zheng, Siyu Wang, Guang Yang, Zihao Zhang, Yu Zhong, Enzhu Gao, Xinhao Zheng, Xueting Wang, Shouming Li, Yunkai Gao, Siming Lan, Mingfei Han, Xing Hu, Dusan Malic, Christian Fruhwirth-Reisinger, Alexander Prutsch, Wei Lin, Samuel Schuller, Horst Possegger, Linfeng Li, Jian Zhao, Zepeng Yang, Yuhang Song, Bojun Lin, Tianle Zhang, Yuchen Yuan, Chi Zhang, Xuelong Li, Youngseok Kim, Sihwan Hwang, Hyeonjun Jeong, Aodi Wu, Xubo Luo, Erjia Xiao, Lingfeng Zhang, Yingbo Tang, Hao Cheng, Renjing Xu, Wenbo Ding, Lei Zhou, Long Chen, Hangjun Ye, Xiaoshuai Hao, Shuangzhi Li, Junlong Shen, Xingyu Li, Hao Ruan, Jinliang Lin, Zhiming Luo, Yu Zang, Cheng Wang, Hanshi Wang, Xijie Gong, Yixiang Yang, Qianli Ma, Zhipeng Zhang, Wenxiang Shi, Jingmeng Zhou, Weijun Zeng, Kexin Xu, Yuchen Zhang, Haoxiang Fu, Ruibin Hu, Yanbiao Ma, Xiyan Feng, Wenbo Zhang, Lu Zhang, Yunzhi Zhuge, Huchuan Lu, You He, Seungjun Yu, Junsung Park, Youngsun Lim, Hyunjung Shim, Faduol Liang, Zihang Wang, Yiming Peng, Guanyu Zong, Xu Li, Binghao Wang, Hao Wei, Yongxin Ma, Yunke Shi, Shuaipeng Liu, Dong Kong, Yongchun Lin, Huitong Yang, Liang Lei, Haoang Li, Xinliang Zhang, Zhiyong Wang, Xiaofeng Wang, Yuxia Fu, Yadan Luo, Djamel Etcheberry, Yang Li, Congfei Li, Yuxiang Sun, Wenkai Zhu, Wang Xu, Linru Li, Longjie Liao, Jun Yan, Benwu Wang, Xueliang Ren, Xiaoyu Yue, Jixian Zheng, Jinfeng Wu, Shurui Qin, Wei Cong, and Yao He. The RoboSense challenge: Sense anything, navigate anywhere, adapt across platforms. <https://robosense2025.github.io>, 2025.
- [12] Lingdong Kong, Shaoyuan Xie, Hanjiang Hu, Yaru Niu, Wei Tsang Ooi, Benoit R. Cottareau, Lai Xing Ng, Yuexin Ma, Wenwei Zhang, Liang Pan, Kai Chen, Ziwei Liu, Weichao Qiu, Wei Zhang, Xu Cao, Hao Lu, Ying-Cong Chen, et al. The RoboDrive challenge: Drive anytime anywhere in any condition. *arXiv preprint arXiv:2405.08816*, 2024.
- [13] Shaoyuan Xie, Lingdong Kong, Wenwei Zhang, Jiawei Ren, Liang Pan, Kai Chen, and Ziwei Liu. RoboBEV: Towards robust bird’s eye view perception under corruptions. *arXiv preprint arXiv:2304.06719*, 2023.
- [14] Xiang Xu et al. Beyond one shot, beyond one perspective: Cross-view and long-horizon distillation for better LiDAR representations. In *IEEE/CVF International Conference on Computer Vision*, pages 25506–25518, 2025.
- [15] Xiaoshuai Hao et al. SafeMap: Robust HD map construction from incomplete observations. In *International Conference on Machine Learning*, pages 22091–22102. PMLR, 2025.
- [16] GLM-V Team. Glm-4.5v and glm-4.1v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning, 2025.
- [17] Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. InternVL3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025.
- [18] Lingdong Kong, Wesley Yang, Jianbiao Mei, Youquan Liu, Ao Liang, Dekai Zhu, Dongyue Lu, Wei Yin, Xiaotao Hu, Mingkai Jia, Junyuan Deng, Kaiwen Zhang, Yang Wu, Tianyi Yan, Shenyuan Gao, Song Wang, Linfeng Li, Liang Pan, Yong Liu, Jianke Zhu, Wei Tsang Ooi, Steven C. H. Hoi, and Ziwei Liu. 3D and 4D world modeling: A survey. *arXiv preprint arXiv:2509.07996*, 2025.
- [19] Tianyi Yan, Tao Tang, Xingtai Gui, Yongkang Li, Jiasen Zhesng, Weiyao Huang, et al. AD-R1: Closed-loop reinforcement learning for end-to-end autonomous driving with impartial world models. *arXiv preprint arXiv:2511.20325*, 2025.
- [20] Shaoyuan Xie et al. Benchmarking and improving bird’s eye view perception robustness in autonomous driving. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(5):3878–3894, 2025.
- [21] Xiaoshuai Hao, Mengchuan Wei, Yifan Yang, et al. Is your HD map constructor reliable under sensor corruptions? In *Advances in Neural Information Processing Systems*, volume 37, pages 22441–22482, 2024.
- [22] Xiaoshuai Hao, Guanqun Liu, Yuting Zhao, et al. MSC-Bench: Benchmarking and analyzing multi-sensor corruption for driving perception. *arXiv preprint arXiv:2501.01037*, 2025.
- [23] Rong Li et al. SeeGround: See and ground for zero-shot open-vocabulary 3D visual grounding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3707–3717, 2025.
- [24] Mohsen Gholami, Ahmad Rezaei, Zhou Weimin, Yong Zhang, and Mohammad Akbari. Spatial reasoning with vision-language models in ego-centric multi-view scenes. *arXiv preprint arXiv:2509.06266*, 2025.
- [25] Murray Shanahan, Kyle McDonell, and Laria Reynolds. Role play with large language models. *Nature*, 623(7987):493–498, 2023.
- [26] Zekun Moore Wang, Zhongyuan Peng, Haoran Que, Jiaheng Liu, Wangchunshu Zhou, Yuhao Wu, Hongcheng Guo, Ruitong Gan, Zehao Ni, Jian Yang, et al. Rolelm: Benchmarking, eliciting, and enhancing role-playing abilities of large language models. *arXiv preprint arXiv:2310.00746*, 2023.

- [27] 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.
- [28] Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. *Advances in neural information processing systems*, 36:11809–11822, 2023.
- [29] Matthew Renze and Erhan Guven. Self-reflection in llm agents: Effects on problem-solving performance. *arXiv preprint arXiv:2405.06682*, 2024.
- [30] Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Tianyu Liu, et al. A survey on in-context learning. *arXiv preprint arXiv:2301.00234*, 2022.
- [31] Sicheng Feng, Kaiwen Tuo, Song Wang, et al. RewardMap: Tackling sparse rewards in fine-grained visual reasoning via multi-stage reinforcement learning. *arXiv preprint arXiv:2510.02240*, 2025.
- [32] Junda Wu, Zhehao Zhang, Yu Xia, Xintong Li, Zhaoyang Xia, Aaron Chang, Tong Yu, Sungchul Kim, Ryan A Rossi, Ruiyi Zhang, et al. Visual prompting in multimodal large language models: A survey. *arXiv preprint arXiv:2409.15310*, 2024.
- [33] Ruochen Wang, Sohyun An, Minhao Cheng, Tianyi Zhou, Sung Ju Hwang, and Cho-Jui Hsieh. One prompt is not enough: Automated construction of a mixture-of-expert prompts. *arXiv preprint arXiv:2407.00256*, 2024.
- [34] Xiang Xu, Lingdong Kong, Hui Shuai, Liang Pan, Ziwei Liu, and Qingshan Liu. LiMoE: Mixture of LiDAR representation learners from automotive scenes. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27368–27379, 2025.
- [35] Lingdong Kong, Yaru Niu, Shaoyuan Xie, Hanjiang Hu, Lai Xing Ng, Benoit Cottureau, Liangjun Zhang, Hesheng Wang, et al. The RoboDepth challenge: Methods and advancements towards robust depth estimation. *arXiv preprint arXiv:2307.15061*, 2023.
- [36] Lingdong Kong, Shaoyuan Xie, Hanjiang Hu, Lai Xing Ng, Benoit R. Cottureau, and Wei Tsang Ooi. RoboDepth: Robust out-of-distribution depth estimation under corruptions. In *Advances in Neural Information Processing Systems*, volume 36, pages 21298–21342, 2023.
- [37] Zeying Gong, Tianshuai Hu, Ronghe Qiu, and Junwei Liang. From cognition to precognition: A future-aware framework for social navigation. In *IEEE International Conference on Robotics and Automation*, pages 9122–9129, 2025.
- [38] Ye Li, Lingdong Kong, Hanjiang Hu, Xiaohao Xu, and Xiaonan Huang. Is your LiDAR placement optimized for 3D scene understanding? In *Advances in Neural Information Processing Systems*, volume 37, pages 34980–35017, 2024.
- [39] Meng Chu, Zhedong Zheng, Wei Ji, Tingyu Wang, and Tat-Seng Chua. Towards natural language-guided drones: GeoText-1652 benchmark with spatial relation matching. In *European Conference on Computer Vision*, pages 213–231, 2024.
- [40] Ao Liang et al. Perspective-invariant 3D object detection. In *IEEE/CVF International Conference on Computer Vision*, pages 27725–27738, 2025.
- [41] Lingdong Kong, Youquan Liu, Xin Li, Runnan Chen, Wenwei Zhang, Jiawei Ren, Liang Pan, Kai Chen, and Ziwei Liu. Robo3D: Towards robust and reliable 3D perception against corruptions. In *IEEE/CVF International Conference on Computer Vision*, pages 19994–20006, 2023.
- [42] Rong Li, Yuhao Dong, Tianshuai Hu, et al. 3EED: Ground everything everywhere in 3D. *arXiv preprint arXiv:2511.01755*, 2025.