Task Aware Prompt Routing and CoT Augmented Fine Tuning for Driving VQA

Hanshi Wang, Xijie Gong, Yixiang Yang, Qianli Ma, Zhipeng Zhang*
AutoLab, School of Artificial Intelligence, Shanghai Jiao Tong University
{hanshi.wang.cv, zhipeng.zhang.cv}@outlook.com

Abstract

We present a unified vision–language framework for RoboSense 2025 Track 1 that integrates perception, prediction, and planning from six surround-view cameras with optional temporal history. Our approach combines training-free, inference-time reasoning with Chain-of-Thought (CoT) augmented fine-tuning of a general-purpose vision–language model. At inference, a lightweight query analyzer performs question-aware prompt routing and dynamically controls both camera-view selection and temporal context. It disables historical frames for corruption identification and short perception MCQs, while enabling strided history for prediction and planning tasks. The analyzer also parses camera tags to filter relevant viewpoints. For model adaptation, we fine-tune Qwen-VL-72B on DriveLM using LoRA, enriching the dataset with concise reasoning traces generated by ChatGLM conditioned on the question and ground-truth answer. Temporal augmentation is applied by adding history frames around t_0 and aligning each instance with the router's temporal policy. On the official RoboSense Track 1 test set, our method achieves a weighted overall score of 64.29 – surpassing the official baseline by 20.59 points – and demonstrates strong robustness, especially under corrupted visual inputs. These results underline the importance of combining task-aware inference routing with structured, CoT-augmented domain adaptation for reliable language-driven autonomous driving.

1. Introduction

Language-guided reasoning has emerged as a powerful tool in autonomous driving, offering a complementary modality that encodes road semantics, interaction patterns, and intent priors [1–3]. When combined with visual perception, language provides a high-level abstraction that improves robustness to distribution shifts, partial observability, and sensor degradation [4–10]. RoboSense 2025 Challenge Track 1, *Driving with Language*, operationalizes this goal by evaluating models across a diverse set of perception, prediction, and planning questions [11]. Systems must process six surround-view images and maintain stable performance even under significant visual corruptions, which are heavily emphasized in the final scoring.

Although modern vision-language models (VLMs) have shown impressive general-purpose reasoning capabilities, they remain unreliable in driving scenarios, particularly under degraded inputs [2, 12–24]. Prior work [11] reveals that VLMs often over-rely on linguistic priors and fail to ground their decisions in visual evidence when images are corrupted or ambiguous. This limitation motivates the need for task-specific guidance during inference and domain adaptation during training.

To bridge this gap, we propose a framework that integrates **training-free inference-time reasoning** with **training-based domain alignment**. At inference time, a query analyzer parses each question and performs task-aware prompt routing. It determines which camera views should be included and whether temporal history is beneficial. History is disabled for corruption identification and short perception MCQs, but enabled with a strided sampling strategy for prediction and planning tasks where temporal cues are critical. Further, CoT reasoning is invoked selectively only when it improves reliability.

On the training side, we adapt Qwen-VL-72B [25] to the autonomous driving domain using supervised instruction fine-tuning with LoRA [26] on DriveLM [1]. Because DriveLM [1] does not include reasoning traces, we augment the data with concise chain-of-thought annotations produced by ChatGLM [27], ensuring the model learns stepwise task logic. We also append temporal history frames around t_0 and align them with the inference router's temporal policy, achieving consistent training–inference behavior.

We evaluate our method on the full RoboSense 2025 Track 1 benchmark, which spans perception multiple-choice questions, descriptive scene understanding, future-motion prediction, and planning decisions. Our approach attains a weighted overall score of 64.29, significantly outperforming the official baseline. The model demonstrates strong reasoning capability, high stability under corrupted conditions, and

*corresponding author

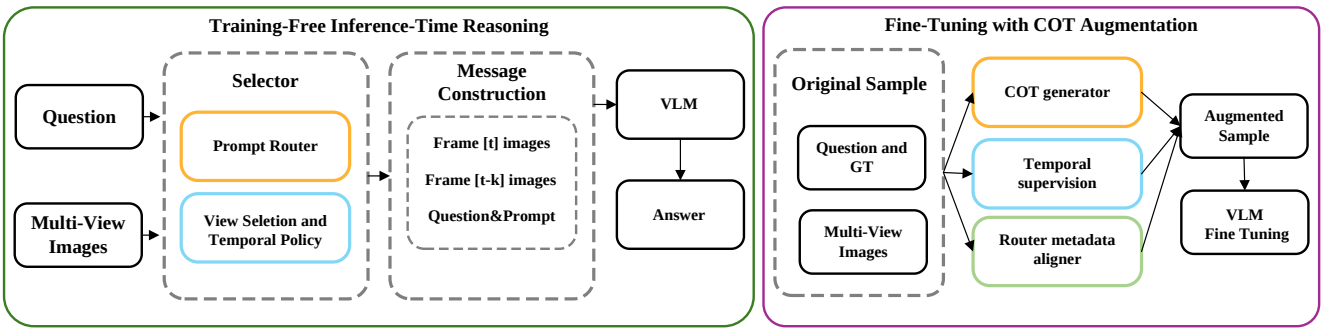

Figure 1. Overview of the proposed framework.

balanced performance across all task types.

In summary, our contributions are as follows: (1) a training-free inference framework featuring task-aware prompt routing, dynamic view selection, and temporal policy control; (2) a domain-adapted VLM trained with LoRA and CoT-augmented instruction tuning to strengthen structured reasoning in autonomous driving.

2. Methodology

2.1. Overview

Our solution is built on Qwen-VL-72B that combines training-free inference with lightweight fine-tuning. At inference time, a question-aware router selects a task-specific prompt and assembles the minimal visual and temporal context required by the query. This yields modular message construction and stable reasoning across tasks. For training, we fine-tune the backbone on DriveLM with concise chains of thought. Overall, the design tightly couples training with inference and delivers robust and consistent performance across diverse driving tasks.

2.2. Training-Free Inference-Time Reasoning

Question Aware Prompt Routing. We maintain a compact library of task-specific system prompts and route each query to its corresponding prompt through pattern-based matching. Unmatched questions fall back to a default template. The router identifies salient keywords and discourse cues in the input and selects the appropriate template. This modular design avoids reliance on a single prompt and enables customized handling of diverse question types.

View Selection and Temporal Policy. Building on the chosen prompt, the router also determines which images enter the context and how temporal cues are used. For corruption MCQs and short perception MCQs, we disable history by design and use only the current frame at t_0. For temporally dependent tasks such as prediction and planning, we include past frames sampled from a history buffer with a fixed stride. Each sampled group is preceded by a short caption that marks its timestamp, for example $[t_0]$ for the current frame and $[t-k]$ for a past step. For description and action questions, we parse camera tags in the question and retain only the referenced views. If an action query omits the front camera, we explicitly include `CAM_FRONT` to preserve forward context. Views follow a fixed camera order. These policies yield reproducible message construction and improved robustness under noisy data loaders.

2.3. Fine-Tuning with COT Augmentation

We fine-tune the backbone on DriveLM. Since the original annotations lack explicit chains of thought, we reannotate the datset with stepwise rationales by prompting ChatGLM to generate concise reasoning traces conditioned on each question and its ground-truth answer so that the explanation remains grounded and self-consistent. In parallel, we enrich temporal supervision by constructing additional history frames around the decision frame t_0 and by aligning each item with the router metadata that governs view selection and temporal policy, which exposes the model to longer motion cues and improves temporal assignment.

3. Experiments

3.1. Dataset

We use the official data provided by the *RoboSense Challenge 2025* [28] held at IROS 2025. This competition builds upon the legacy of the *RoboDepth Challenge 2023* [29, 30] at ICRA 2023 and the *RoboDrive Challenge 2024* [31, 32] 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 [11, 33–38]. Specifically, this task is built upon the **DriveBench** dataset [11] in **Track 1**, which evaluates vision-language models in autonomous driving through perception, prediction, and planning questions under both clean and corrupted visual conditions.

3.2. Implementation Details

The input uses six surround cameras in a fixed canonical order with camera tags embedded in the prompt. History frames are included only when the selected template benefits from motion cues. We treat t_0 as the decision frame and use preceding frames to estimate heading changes. We fine-tune Qwen2.5-VL-72B-Instruct [25] with supervised instruction tuning and a LoRA configuration with rank 8 and scaling factor 16, applying adapters across all target modules while keeping the vision tower and the multimodal projector frozen. Optimization uses AdamW with a learning rate of 2×10^{-5} and a cosine schedule for 3 epochs with 20 warmup steps.

3.3. Main Experiment

Tab. 1 summarizes Phase 2 test results where image is corrupted. Our model raises the aggregate score from 43.7 to 64.29 for a gain of +20.59. The largest improvements occur in perception VQA for scenes +40.12 and objects +28.69, which we attribute to question-aware prompt routing and grounded view selection under corrupted inputs. Prediction accuracy increases by +11.45 due to a temporal policy that samples history only when motion cues are required. Corruption MCQ improves by +17.34, and planning VQA gains +22.93 on object questions and +27.73 on scene questions, enabled by selective chain-of-thought prompting and effective fine-tuning. Together, these results demonstrate strong robustness across diverse corruption types.

Table 1. Phase 2 results on the Track 1 test split where all inputs are corrupted. MCQ entries are accurate %. VQA entries are weighted VQA scores %. The overall score follows the official weighting.

Task	Baseline	Ours	Gain
Perception MCQ	78.60	95.92	+17.32
Perception VQA Object	21.70	50.39	+28.69
Perception VQA Scene	19.30	59.42	+40.12
Prediction MCQ	61.60	73.05	+11.45
Corruption MCQ	81.70	99.04	+17.34
Planning VQA Scene	31.30	59.03	+27.73
Planning VQA Object	30.80	53.73	+22.93
Overall	43.70	**64.29**	**+20.59**

4. Conclusion

We introduce a robust vision language system for multi-view driving questions that couples training-free inference with lightweight fine-tuning. A question-aware router composes structured prompts and assembles visual and temporal context. We also fine-tune on DriveLM with concise chain of thought augmentation. On RoboSense 2025 Track 1, the system delivers large gains in perception and prediction and stable planning under heavy corruptions. Future work will add explicit COT reasoning, and cross-view scene graphs, and will extend evaluation to closed-loop settings.

References

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

[2] Jiahan Li, Zhiqi Li, and Tong Lu. Driving with InternVL: Oustanding champion in the track on driving with language of the autonomous grand challenge at CVPR 2024. *arXiv preprint arXiv:2412.07247*, 2024.

[3] 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.

[4] 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.

[5] Xiaoshuai Hao et al. SafeMap: Robust HD map construction from incomplete observations. In *International Conference on Machine Learning*, pages 22091–22102. PMLR, 2025.

[6] 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.

[7] Lingdong Kong, Youquan Liu, Runnan Chen, Yuexin Ma, Xinge Zhu, Yikang Li, Yuenan Hou, Yu Qiao, and Ziwei Liu. Rethinking range view representation for LiDAR segmentation. In *IEEE/CVF International Conference on Computer Vision*, pages 228–240, 2023.

[8] 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.

[9] Xuzhi Wang, Xinran Wu, Song Wang, et al. Monocular semantic scene completion via masked recurrent networks. In *IEEE/CVF International Conference on Computer Vision*, pages 24811–24822, 2025.

[10] 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.

[11] 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.

[12] 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.

[13] Ye Li et al. Optimizing LiDAR placements for robust driving perception in adverse conditions. *arXiv preprint arXiv:2403.17009*, 2024.

[14] 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.

[15] Youquan Liu et al. Multi-space alignments towards universal LiDAR segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14648–14661, 2024.

[16] Lingdong Kong, Niamul Quader, and Venice Erin Liong. ConDA: Unsupervised domain adaptation for LiDAR segmentation via regularized domain concatenation. In *IEEE International Conference on Robotics and Automation*, pages 9338–9345, 2023.

[17] 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.

[18] Xidong Peng, Runnan Chen, Feng Qiao, et al. Learning to adapt SAM for segmenting cross-domain point clouds. In *European Conference on Computer Vision*, pages 54–71. Springer, 2024.

[19] Hengwei Bian et al. DynamicCity: Large-scale 4D occupancy generation from dynamic scenes. In *International Conference on Learning Representations*, 2025.

[20] Xiang Xu et al. 4D contrastive superflows are dense 3D representation learners. In *European Conference on Computer Vision*, pages 58–80, 2024.

[21] Xiang Xu et al. FRNet: Frustum-range networks for scalable LiDAR segmentation. *IEEE Transactions on Image Processing*, 34:2173–2186, 2025.

[22] Lingdong Kong et al. LargeAD: Large-scale cross-sensor data pretraining for autonomous driving. *arXiv preprint arXiv:2501.04005*, 2025.

[23] 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.

[24] 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.

[25] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

[26] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. LoRA: Low-rank adaptation of large language models. *International Conference on Learning Representations*, 2022.

[27] GLM-V Team. Glm-4.5v and glm-4.1v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning, 2025.

[28] Lingdong Kong, Shaoyuan Xie, Zeying 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. Cottereau, 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, Xinhan Zheng, Xueting Wang, Shouming Li, Yunkai Gao, Siming Lan, Mingfei Han, Xing Hu, Dusan Malic, Christian Fruhwirth-Reisinger, Alexander Prutsch, Wei Lin, Samuel Schulter, 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, Faduo 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, Djamahl Etchegaray, 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.

[29] Lingdong Kong, Yaru Niu, Shaoyuan Xie, Hanjiang Hu, Lai Xing Ng, Benoit Cottereau, Liangjun Zhang, Hesheng Wang, et al. The RoboDepth challenge: Methods and advancements towards robust depth estimation. *arXiv preprint arXiv:2307.15061*, 2023.

[30] Lingdong Kong, Shaoyuan Xie, Hanjiang Hu, Lai Xing Ng, Benoit R. Cottereau, 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.

[31] Lingdong Kong, Shaoyuan Xie, Hanjiang Hu, Yaru Niu, Wei Tsang Ooi, Benoit R. Cottereau, 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.

[32] 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.

[33] 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.

[34] 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.

[35] 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.

[36] Ao Liang et al. Perspective-invariant 3D object detection. In *IEEE/CVF International Conference on Computer Vision*, pages 27725–27738, 2025.

[37] 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.

[38] Rong Li, Yuhao Dong, Tianshuai Hu, et al. 3EED: Ground everything everywhere in 3D. *arXiv preprint arXiv:2511.01755*, 2025.