Enhancing Multi-View Driving VLMs via Pseudo-Label Pretraining and Long-Tail Balancing

Jiangpeng Zheng
Tianjin University of Technology
zjp_1997@stud.tjut.edu.cn

Ji Ao
aoji.vip@gmail.com

Guang Yang
Tianjin University of Technology
yangguang@stud.tjut.edu.cn

Siyu Wang
1410813444@qq.com

Abstract

Vision–language models (VLMs) for autonomous driving face critical challenges when dealing with multi-view camera inputs and degraded visual conditions, including limited robustness to corruption, insufficient capability for multi-task integration, and constrained accuracy and adaptability. To address these issues, we propose a two-stage optimization framework built upon InternVL3-8B. In the first stage, pseudo-label pretraining with chain-of-thought (CoT) reasoning and fixed-sequence multi-view image concatenation are employed to enhance input understanding. A large-scale training set derived from the DriveLM benchmark is further constructed, covering four major tasks – perception, prediction, planning, and behavior – to strengthen multi-view semantic understanding and reasoning. In the second stage, the CoT-free model from stage one is refined through three key strategies: targeted handling of long-tail data, hybrid fine-tuning with official and synthetic datasets, and ensemble-based integration. In addition, the task distribution is rebalanced, incorporating a new corruption task to achieve a uniform ratio across all tasks. Experimental results demonstrate that the proposed framework improves the robustness and generalization of VLMs for autonomous driving, outperforming representative baselines such as Qwen2.5-VL-7B. Ablation studies further confirm the effectiveness of multi-view concatenation, pseudo-label pretraining, and long-tail balancing in enhancing overall performance.

1. Introduction

Autonomous driving is steadily advancing toward richer multimodal interaction, where Vision-Language Models (VLMs) play an increasingly central role in connecting human-level instructions with machine-level driving decisions [1–5]. By jointly modeling natural language and visual perception, VLMs offer a unified framework for describing complex driving scenarios, interpreting human intent, and enabling interpretable decision-making [6–14]. The *"Drive with Language"* challenge [15] further formalizes this objective: participating models must handle perception [16–21], prediction [22], and planning [23–25] questions under multi-view camera inputs, while maintaining stable performance under a range of visually degraded conditions [26–31]. This setup aligns with the multi-task collaborative reasoning emphasized in *DriveLM: Driving with Graph Visual Question Answering* [22], and simultaneously exposes structural limitations in current VLM-based driving systems.

While recent advancements have improved VLM performance in driving environments, several challenges remain [22, 32–35]. For instance, *Driving with InternVL* [36] demonstrates that multi-view fusion can strengthen scene understanding, and *Precise Drive with VLM* [37], the winning solution of the PRCV 2024 Autonomous Driving LLM Challenge, shows that task-specific optimization can further enhance performance. However, three critical issues persist.

First, **robustness under visual degradation** is still inadequate. According to *Are VLMs Ready for Autonomous Driving?* [32], mainstream VLMs experience more than a 40% performance drop under corrupted imagery, indicating strong vulnerability to real-world visual variation.

Second, **multi-task integration remains weak**. Many VLMs process perception, prediction, and planning as independent subtasks, lacking the structured, interdependent reasoning that frameworks like DriveLM aim to provide. This fragmented processing hinders consistent cross-task

logic.

Third, there is a persistent **gap between model scale, accuracy, and deployability** [36, 38, 39]. Although large models such as InternVL3-14B exhibit strong potential, their improvements do not scale efficiently with computational overhead, especially when the model is not explicitly aligned with task-specific supervision or balanced data distributions.

To address these limitations, we propose a two-phase optimization framework built upon InternVL3-8B. In the **first phase**, we enhance input-level understanding using two strategies: (1) pseudo-annotation pre-training incorporating Chain-of-Thought (CoT) reasoning to inject structured semantic cues, and (2) fixed-order concatenation of multi-view images to provide the model with consistent spatial grounding across six camera views. We additionally construct a large-scale, diversified training set derived from DriveLM, covering four critical tasks – perception, prediction, planning, and behavior – to improve multi-view semantic representation and cross-task reasoning.

In the **second phase**, we refine the CoT-free model from phase one through three complementary techniques: (1) targeted rebalancing of long-tailed task distributions, (2) hybrid fine-tuning combining official and synthetic datasets to expand visual and linguistic diversity, and (3) multi-scale model ensembling to stabilize predictions and enhance overall robustness under corruption. This phased strategy draws upon lessons from *Precise Drive with VLM*'s multi-view optimization techniques while streamlining inefficiencies noted in earlier versions of *Driving with InternVL*.

Through this structured optimization pipeline, our final model achieves strong multi-task reasoning ability, improved corruption robustness, and consistent performance across all task types in the RoboSense 2025 Track 1 challenge. The results underscore the importance of balancing multi-view fusion, structured supervision, and distribution-aware fine-tuning when adapting VLMs for real-world autonomous driving.

2. Methodology

2.1. Dataset

2.1.1 Phase 1

We adopt DriveLM [22], one of the most representative benchmark datasets, as the foundation for our training dataset. The Phase 1 training dataset encompasses four tasks: perception, prediction, planning, and behavior, each formulated through distinct question types, including multiple-choice questions (MCQs) and visual question answering (VQA). As summarized in Fig. 1, the distribution is: perception 162,480 items (43.0%), planning 123,436 (32.7%), prediction 87,968 (23.3%), and behavior 4,072 (1.1%), for a total of 377,956 QA pairs. This composition maintains perception as the dominant component while ensuring substantial coverage of planning and prediction and a smaller, focused behavior subset – providing a balanced and reliable foundation for model training and evaluation.

We performed the following operations on this dataset: Firstly, we employ InternVL3-8B-Instruct to generate pseudo image captions for all single images. Secondly, we concatenate the six-view images of each sample in sequence and integrate their corresponding captions. Finally, we leverage InternVL3-14B-Instruct to generate chain-of-thought (CoT) pseudo annotations by providing the concatenated images along with the corresponding questions and answers. In the training data, CoT reasoning is denoted by <think>...</think>, while the final answers are denoted by <answer>...</answer>.

2.1.2 Phase 2

We identify two major limitations in the original dataset: (i) insufficient diversity of question types, which fails to adequately cover diverse traffic scenarios; and (ii) imbalanced answer distributions within task categories, which introduces significant skew. These dual deficiencies lead to systematic bias – on the one hand, the lack of exposure to diverse traffic conditions constrains the model's generalization capability; on the other hand, skewed distributions cause the model to memorize frequent patterns and common cases rather than learning their underlying semantics. As a result, model evaluation becomes distorted, and generalization to complex driving environments is hindered.

To address these issues, we performed targeted expansion and refinement of the dataset in Phase 2, with details illustrated in Fig. 2 and summarized in Table 1. On this basis, we further applied a balancing strategy (denoted as total_balance), through which the distribution of the four task types – Perception, Prediction, Planning, and Corruption – was resampled to a 1:1:1:1 ratio. This adjustment substantially enhances both the completeness and fairness of task coverage. In addition, Fig. 3 provides a direct comparison between the original and the expanded question sets, clearly demonstrating the improvements in diversity and balance introduced in Phase 2.

- **Perception**: To strengthen the model's fundamental perception capability, we expand the perception tasks from 400 to 30,529 samples by incorporating object information beyond the original ground-truth annotations. This extension enriches semantic coverage across global scene recognition, local object attributes, and motion status, strengthening the model's perception of complex driving environments.
- **Planning**: We enlarge planning tasks from 600 to 35,444 samples by constructing problems from nuScenes trajectories and scene configurations. The new data emphasize behavior reasoning, collision risk anticipation, and safe

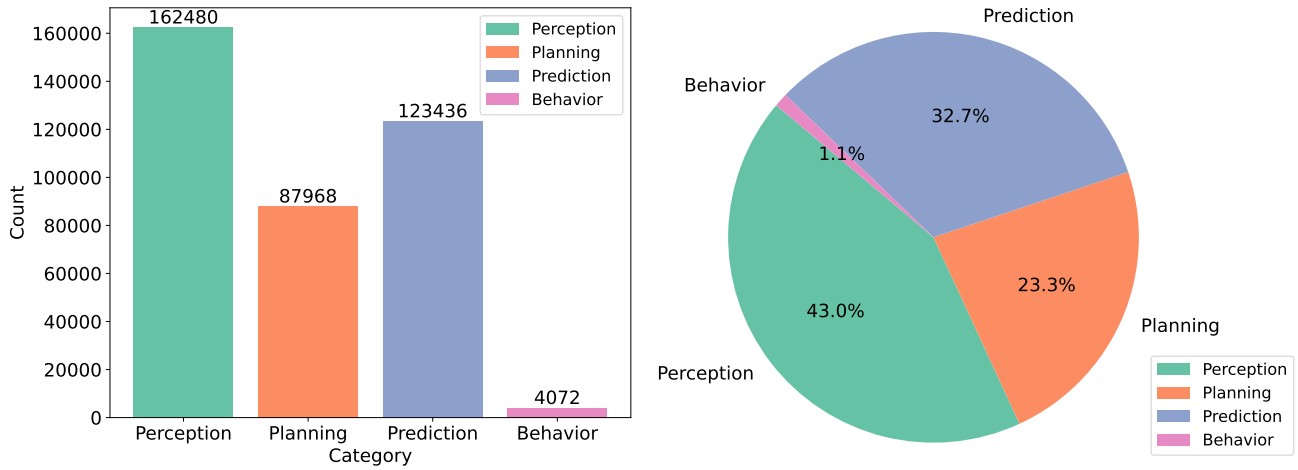

Figure 1. Phase 1: Question Distribution of Perception, Prediction, Planning, and Behavior.

Table 1. Number of Newly Added Questions and Examples

Data Naming	Task Type	# QA	Questions
qa_perception	Perception	30529	1. What is the visual description of <OBJ>? 2. What are the important objects in the current scene? Those objects will be considered for future reasoning and driving decisions. 3. What is the moving status of object <OBJ>? Please select the correct answer from the following options: A. ... B. ... C. ... D. ...
qa_planning	Planning	35444	1. What actions could the ego vehicle take based on ? Why take this action and what's the probability? 2. What actions taken by the ego vehicle can lead to a collision with <OBJ>? 3. In this scenario, what are safe actions to take for the ego vehicle?
qa_prediction	Prediction	37648	1. What object should the ego vehicle notice first when the ego vehicle is getting to the next possible location? ... 2. Would <OBJ>be in the moving direction of the ego vehicle? 3. Will <OBJ>be in the moving direction of <OBJ>? 4. Will <OBJ>change its motion state based on <OBJ>?

action identification, substantially enhancing the model's proactive decision-making capacity.

- **Prediction**: We scale prediction tasks from 261 to 37,648 samples based on nuScenes trajectories and scene topology. The added questions address motion conflicts, agent interactions, and attention prioritization in sequential decisions, improving the model's ability to anticipate dynamic scene evolution.

2.2. Training

Figure 4 illustrates the overall two-phase training pipeline of our framework. In Phase 1, the model is pre-trained with pseudo-labels, multi-view image concatenation, and optional chain-of-thought (CoT) supervision to enhance input understanding and reasoning ability. In Phase 2, the framework further emphasizes robustness and generalization through balanced data distribution, synthetic augmentation, and ensemble integration. Together, the two stages form a progressive optimization process that systematically improves both perception and decision-making under multi-view and corrupted driving conditions.

2.2.1 Phase 1

Phase 1 aimed to inject driving scene knowledge into the model through pre-training. The process consisted of three steps:

- **Initial Pre-training:** InternVL3-8B-Instruct was employed to generate pseudo captions for all images. InternVL3-8B was then initially pre-trained on this data, yielding Model-a.
- **Multi-view Adaptation:** The six-view images of each sample were concatenated and fed into Model-a, along with their corresponding textual descriptions, for a second round of pre-training. This step enhanced the model's adaptation to panoramic image inputs, resulting in the improved Model-b.
- **Chain-of-Thought Integration:** Based on Model-b, we trained two variants using the official dataset: one with CoT annotations (phase1-model-w-cot) and one without (phase1-model-wo-cot). Of these, the model trained with CoT (phase1-model-w-cot) was selected, and its results were submitted to the Phase 1 leaderboard.

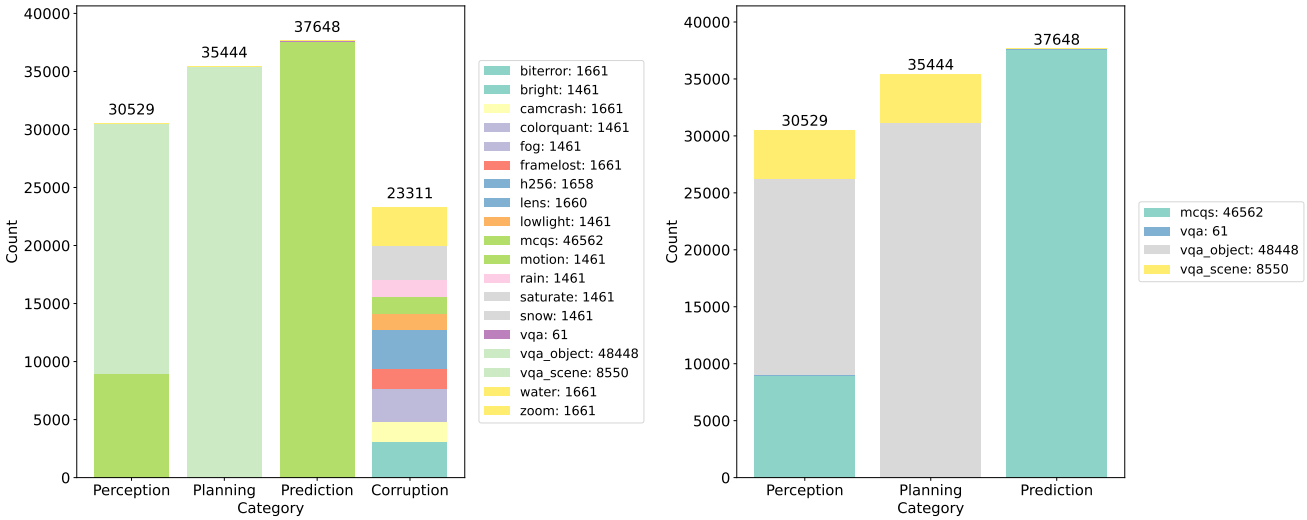

Figure 2. Phase 2: Question Distribution of Perception, Prediction, and Planning

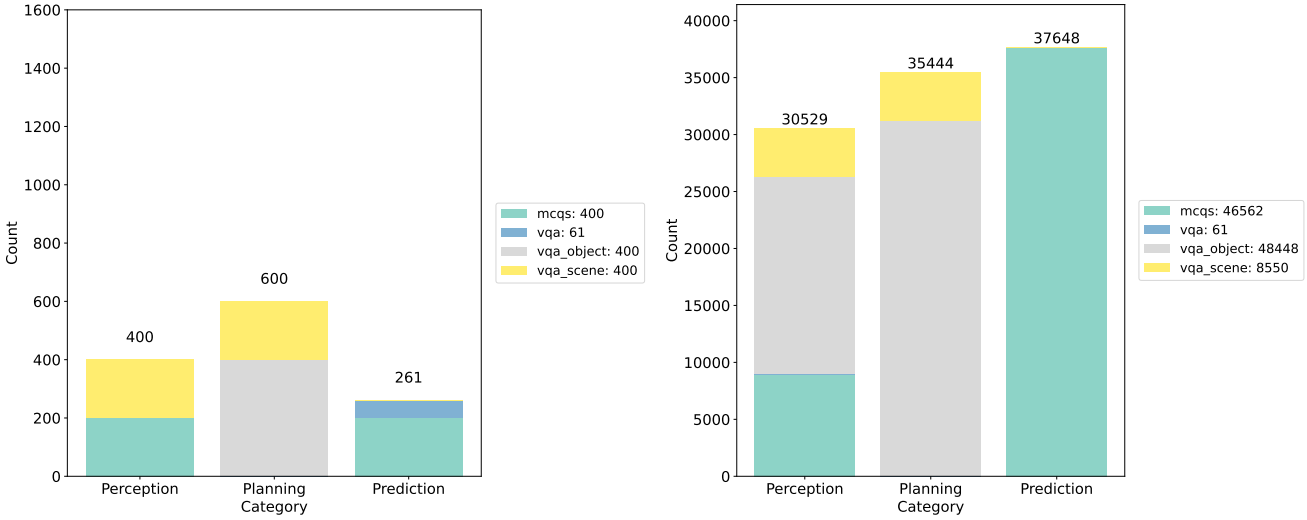

Figure 3. Number of Question Types Before and After Dataset Expansion. The left image shows the original data, while the right image shows the enhanced data.

2.2.2 Input Representation and Preprocessing

To overcome the limited field-of-view issues inherent in single-view images (e.g., missing rear-side obstacles or cross-traffic at intersections), this study adopted a multi-view image fusion strategy. Specifically, images captured by the six core cameras (Front Left, Front, Front Right, Rear Left, Rear, Rear Right) were concatenated according to their spatial logic to form panoramic six-view images as model input. Meanwhile, we designate the CAM_FRONT_LEFT, CAM_FRONT, CAM_FRONT_RIGHT, CAM_BACK_LEFT, CAM_BACK, CAM_BACK_RIGHT, <think>, </think>, <answer>, </answer> as special tokens during training and inference. This approach ensures 360° blind-spot-free environmental information coverage, providing a comprehensive and reliable data foundation for downstream perception, prediction, and planning tasks.

To guide the model in efficiently parsing this specific input format, we uniformly optimized the System Prompt across all training phases, with the core content as follows: *You are a highly trained autonomous driving AI system, capable of interpreting multimodal sensory input (camera images, LiDAR) and natural language instructions. You receive an image that consists of six surrounding camera views. The layout is as follows: The first row contains three images: the images of the left front, front, and right front of the car,*

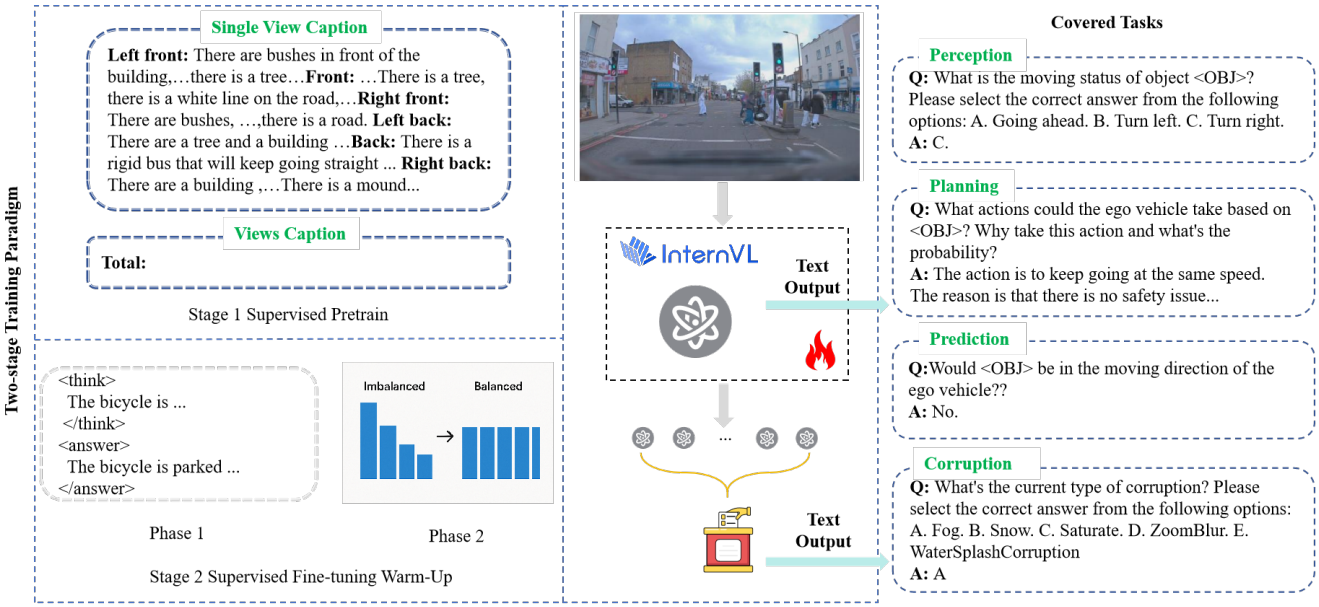

Figure 4. Number of Question Types Before and After Dataset Expansion. The left image shows the original data, while the right image shows the enhanced data. In Phase 1, pseudo-labels from InternVL3-8B are used for initial pre-training; six-view images are spatially concatenated for multi-view pre-training; and CoT data generated by InternVL3-14B-Instruct is leveraged to fine-tune a reasoning-capable model for leaderboard submission. In Phase 2, the non-CoT variant from Phase 1 is further fine-tuned on balanced, multi-view SFT data to enhance robustness, with final predictions obtained via category-wise voting over an ensemble of models.

Table 2. Overall results

Phase	Perception MCQs	Perception VQA-Object	Perception VQA-Scene	Prediction MCQs	Planning VQA-Object	Planning VQA-Scene	Corruption MCQs	Avg
Phase 1	92.45	60.49	47.19	68.77	52.40	61.17	-	63.39
Phase 2	61.22	57.74	50.93	69.92	49.52	49.18	100.00	60.69

respectively. The second row contains three images: the images of the left back, back, and right back of the car, respectively. The object coordinates are provided in the format of <id, camera_view, x, y>. Your task is to provide safe, reasonable, and accurate judgment in various driving scenarios.

2.2.3 Phase 2

Phase 2 focused on further enhancing model performance via Instruction Fine-Tuning (SFT). Since experiments in Phase 1 indicated that CoT annotations did not yield significant performance gains, phase1-model-wo-cot was selected as the pre-trained model for this phase. Regarding training data, we similarly used concatenated six-view images and applied special handling to long-tailed data to construct a high-quality SFT dataset. Building on this, we fine-tuned the model using a mixture of official and synthetic data, and explored ensemble strategies with models of different scales. The final predictions were obtained by performing category-wise voting on the test set. It is noteworthy that no CoT data were used during the SFT in this phase.

2.3. Inference

During the inference stage, we follow the same image processing procedure as in training, concatenating the six-view images of each sample in sequence and feeding them to the model together with the corresponding question. For CoT-based models, we extract the content enclosed within <answer>...</answer> as the final result.

3. Experiments

3.1. Dataset

We use the official data provided by the *RoboSense Challenge 2025* [40] held at IROS 2025. This competition builds upon the legacy of the *RoboDepth Challenge 2023* [41, 42] at ICRA 2023 and the *RoboDrive Challenge 2024* [43, 44] at ICRA 2024, continuing the collective effort to advance robust and scalable robot perception. Each track in this compe-

Table 3. Phase 1 comparative study of different VLMs

Pre-trained Model	Perception MCQs	Perception VQA-Object	Perception VQA-Scene	Prediction MCQs	Planning VQA-Object	Planning VQA-Scene	Avg
Qwen2.5VL-7B	47.17	42.75	28.98	59.00	32.57	46.46	42.80
InternVL3-8B	**86.79**	46.48	39.06	**65.52**	**51.68**	55.68	**57.69**
InternVL3-38B	83.02	**49.55**	**47.19**	62.45	50.90	**57.38**	57.64

Table 4. Phase 2 comparative study of different pre-trained models

Pre-trained Model	Perception MCQs	Perception VQA-Object	Perception VQA-Scene	Prediction MCQs	Planning VQA-Object	Planning VQA-Scene	Corruption MCQs	Avg
InternVL3-8B	44.90	29.80	46.63	61.79	41.92	47.06	100.00	50.87
phase1-model-w-cot	39.80	46.93	45.45	61.21	45.17	49.28	93.27	53.98
phase1-model-wo-cot	**44.90**	**48.74**	**48.86**	**65.62**	**53.79**	**51.49**	**100.00**	**57.88**

tition is grounded on an established benchmark designed for evaluating real-world robustness and generalization [32, 45–48]. Specifically, this task is built upon the **DriveBench** dataset [32] 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. Experimental Setups

Our experiments are primarily conducted on InternVL3-8B, without altering the overall network architecture. In both training phases, we mainly rely on the official dataset, while making slight adjustments to the sample ratios. During evaluation, we follow the official testing pipeline and use the provided codebase.

3.3. Implementation Details

During both the pre-training and supervised fine-tuning stages, the Vision Transformer (ViT) backbone was frozen, while the MLP adapter and the Large Language Model (LLM) components were trained. All models were trained for only one epoch with a fixed learning rate of 2e-5 and a global batch size of 128. The experiments were performed on a single node with 8*A100 GPUs running Ubuntu, using PyTorch with DeepSpeed (ZeRO-2). For inference, we set the temperature to 0 and top-p to 0.7 to ensure the accuracy and stability of the outputs.

3.4. Overall Results

As shown in Table 2, our approach achieved competitive overall performance in both phases. The Phase 1 model attained an average score of 63.39, demonstrating particularly outstanding performance on the Perception MCQs task with a high score of 92.45. The Phase 2 model achieved an average score of 60.69. While its scores on some tasks were lower than those in Phase 1, it achieved a perfect score of 100.0 on the newly added Corruption MCQs task, proving the model's robustness in handling image corruptions. The overall results indicate that our method delivers stable and effective performance across various subtasks.

3.5. Comparative Study

3.5.1 Phase 1

As shown in Table 3, among the model selections in Phase 1, the InternVL3 series models significantly outperformed Qwen2.5-VL-7B. Specifically, the InternVL3-8B model, which served as our final baseline, achieved an average score (57.69) substantially higher than that of Qwen2.5-VL-7B (42.80), representing an absolute advantage of 14.89 points. Notably, on the Perception MCQs task, InternVL3-8B (86.79) outperformed Qwen2.5-VL-7B (47.17) by nearly 40 points, showcasing its superior capability in visual information processing. Consequently, InternVL3-8B was selected as the baseline model for Phase 1.

3.5.2 Phase 2

In Phase 2, we compared the effects of different pre-trained models. Table 4 shows that the phase1-model-wo-cot model, derived from Phase 1 training, performed the best. Its average score (57.88) was significantly higher than that of the model fine-tuned directly from the original InternVL3-8B (50.87) and the phase1-model-w-cot model that utilized Chain-of-Thought (53.98). This indicates that the pre-training process in Phase 1 successfully infused the model with crucial driving scene knowledge, laying a solid foundation for Phase 2 tasks, while CoT data did not provide additional gains in this stage.

Table 5. Phase 1 ablation study of strategies on InternVL3-8B

Ablation Setting			Evaluation Metrics (%)						
Image Concat.	Caption Pre-train	CoT Supervision	Perception MCQs	Perception VQA-Object	Perception VQA-Scene	Prediction MCQs	Planning VQA-Object	Planning VQA-Scene	Avg
×	×	×	86.79	46.48	39.06	65.52	51.68	55.68	57.69
✓	×	×	41.51	59.82	50.31	69.16	52.75	59.36	61.22
✓	✓	×	81.13	61.25	44.69	68.20	53.50	60.75	62.79
✓	✓	✓	92.45	60.49	47.19	68.77	52.40	61.17	63.39

Table 6. Phase 2 ablation study of different datasets on phase1-model-wo-cot

datasets	Perception MCQs	Perception VQA-Object	Perception VQA-Scene	Prediction MCQs	Planning VQA-Object	Planning VQA-Scene	Corruption MCQs	Avg
official phase2	44.90	48.74	48.86	65.62	53.79	51.49	100.00	57.88
+ qa_perception	41.84	56.12	50.10	67.60	47.94	49.89	100.00	58.81
+ qa_planning	43.88	55.74	48.56	66.90	49.01	50.11	100.00	58.67
+ qa_prediction	42.86	56.04	47.63	67.02	46.06	50.34	100.00	58.46
+ total	43.88	56.24	47.87	68.29	49.96	49.89	95.19	59.05
+ total_balance	42.86	57.08	48.21	68.18	51.90	50.40	100.00	59.65

3.6. Ablation Study

3.6.1 Phase 1

We conducted a systematic ablation study on the core training strategies in Phase 1, with results presented in Table 5. Baseline Model (No concatenation, No pre-training): Achieved an average score of 57.69. Introducing Six-View Image Concatenation: The average score increased to 61.22 (+3.53), with a particularly significant improvement on the Perception VQA-Object task (from 46.48 to 59.82), which requires spatial understanding, demonstrating the effectiveness of the multi-view input. Adding Caption-Based Pre-training: After incorporating pre-training on top of concatenation, the average score further increased to 62.79 (+1.57), consolidating the model's performance across multiple tasks. Introducing Chain-of-Thought Supervision: The full model, integrating all strategies, achieved the highest average score of 63.39. The addition of CoT pushed the performance on Perception MCQs to its peak (92.45), indicating that complex reasoning supervision aids in handling challenging multiple-choice questions.

3.6.2 Phase 2

The ablation study in Phase 2 focused on data configurations, as shown in 6. Starting from the official Phase 2 data as the baseline (Avg: 57.88), we incrementally added synthetic datasets qa_perception, qa_planning, and qa_prediction. Adding any single category of synthetic data (+ qa_perception, + qa_planning, + qa_prediction) consistently improved the average score to 58.81, 58.67, and 58.46, respectively, demonstrating the general beneficial effect of supplementary data. We combined these three parts of synthetic data called *total*, and further improved the average score to 59.05 points. After applying long-tail balancing to the total data, the best average score of 59.65 was achieved, which is 1.77 points higher than the baseline. This strategy notably enhanced performance on tasks such as Perception VQA-Object (57.08) and Planning VQA-Object (51.9), validating the importance of optimizing for data distribution.

4. Conclusion

This work addresses the limitations of vision–language models (VLMs) in autonomous driving, including insufficient robustness to visual degradations, weak multi-task coordination, and limited accuracy and adaptability. We propose a two-stage optimization framework built upon InternVL3-8B, following a progressive design of pretraining for input understanding and fine-tuning for robustness enhancement. In Phase 1, the DriveLM benchmark is employed to ensure comprehensive task coverage, while in Phase 2, perception tasks are substantially expanded (from 400 to 30,529 samples) and task distributions are rebalanced to 1:1:1:1, effectively mitigating the issues of task scarcity and imbalance in the original dataset.

Experimental results demonstrate that the proposed framework significantly improves performance under multi-view and degraded conditions, enhancing adaptability in complex driving scenarios and providing a practical path for optimizing VLMs in autonomous driving. Future work will explore model compression and acceleration techniques to balance accuracy and computational cost, as well as the integration

of large-scale real-world traffic data to further strengthen generalization and practical applicability.

References

[1] 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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11621–11631, 2020.

[2] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the KITTI vision benchmark suite. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3354–3361, 2012.

[3] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, et al. Scalability in perception for autonomous driving: Waymo open dataset. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2446–2454, 2020.

[4] 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.

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

[6] 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.

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

[8] 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.

[9] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *NeurIPS*, 2023.

[10] Runnan Chen et al. CLIP2Scene: Towards label-efficient 3D scene understanding by CLIP. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7020–7030, 2023.

[11] 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.

[12] Youquan Liu et al. Segment any point cloud sequences by distilling vision foundation models. In *Advances in Neural Information Processing Systems*, volume 36, pages 37193–37229, 2023.

[13] 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.

[14] 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.

[15] RoboSense Challenge 2025 Organizers. Robosense challenge 2025: Track 1 - driving with language. https://robosense2025.github.io/track1, 2025.

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

[17] Xuzhi Wang et al. NUC-Net: Non-uniform cylindrical partition network for efficient LiDAR semantic segmentation. *IEEE Transactions on Circuits and Systems for Video Technology*, 35(9):9090–9104, 2025.

[18] Youquan Liu et al. UniSeg: A unified multi-modal LiDAR segmentation network and the openpcseg codebase. In *IEEE/CVF International Conference on Computer Vision*, pages 21662–21673, 2023.

[19] 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.

[20] 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.

[21] Lingdong Kong, Jiawei Ren, Liang Pan, and Ziwei Liu. Lasermix for semi-supervised LiDAR semantic segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21705–21715, 2023.

[22] 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*, pages 256–274. Springer, 2024.

[23] 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.

[24] 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.

[25] 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.

[26] 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.

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

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

[29] 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.

[30] 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.

[31] 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.

[32] 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.

[33] 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.

[34] 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.

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

[36] 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.

[37] Bin Huang, Siyu Wang, Yuanpeng Chen, Yidan Wu, Hui Song, Zifan Ding, Jing Leng, Chengpeng Liang, Peng Xue, Junliang Zhang, et al. Precise drive with VLM: First prize solution for PRCV 2024 Drive LM challenge. *arXiv preprint arXiv:2411.02999*, 2024.

[38] Dongping Chen et al. MLLM-as-a-judge: Assessing multimodal LLM-as-a-judge with vision-language benchmark. In *International Conference on Machine Learning*, 2024.

[39] Seonho Lee, Jiho Choi, Inha Kang, Jiwook Kim, Junsung Park, and Hyunjung Shim. 3D-aware vision-language models fine-tuning with geometric distillation. *arXiv preprint arXiv:2506.09883*, 2025.

[40] 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.

[41] 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.

[42] 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.

[43] 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.

[44] 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.

[45] 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.

[46] 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.

[47] 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.

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