Driving Robustly through Corruptions: Multi-Source LoRA Fine-Tuning of Driving VLMs for Multi-View Reasoning

Yuxia Fu
The University of Queensland
Brisbane, Australia
yuxia.fu@uq.edu.au

Djamahl Etchegaray
The University of Queensland
Brisbane, Australia
uqdetche@uq.edu.au

Yadan Luo
The University of Queensland
Brisbane, Australia
y.luo@uq.edu.au

Abstract

Autonomous driving requires not only a deep understanding of complex real-world environments but also robustness to corrupted visual inputs in order to ensure safe and interpretable decisions. Recent advances in vision-language models (VLMs) offer promising capabilities in scene understanding, reasoning, and instruction following, making them well-suited for high-level decision-making tasks in driving scenarios. In this work, we build upon Senna-VLM, a VLM architecture specifically designed for autonomous driving, which features a Driving Vision Adapter for multi-view image compression and a prompting scheme for spatial context awareness. To further enhance the model's robustness to corrupted inputs and improve its generalization to multiple-choice and open-ended questions, we fine-tune Senna-VLM using a mixture of high-quality QA datasets, including DriveLM and DriveBench. Our solution achieves a top-5 performance in Track 1 of the RoboSense Challenge 2025, highlighting the effectiveness of our approach in perception, prediction, and planning tasks under diverse visual conditions.

1. Introduction

Ensuring reliable and accurate decision-making is a central objective of autonomous driving [1–3]. Modern autonomous systems are expected to perceive complex environments, reason about the future motion of surrounding agents, and generate driving plans that are not only safe but also logically consistent with the driving context [4–7]. However, robust performance under imperfect conditions remains a major challenge for autonomous driving systems [8].

In practice, visual inputs are frequently affected by corruptions such as brightness changes, rain, blur, or partial occlusions [9–14]. These corrupt inputs can significantly degrade perception and reasoning quality, which in turn undermines the reliability of downstream perception, prediction, and planning [15–21].

In parallel, the emergence of vision-language models (VLMs) has opened new avenues for scene understanding [22–27], reasoning [28–32], and instruction-following [33–39] in autonomous driving. By aligning visual perception with language-based reasoning, VLMs exhibit strong performance in tasks such as open-ended question answering, visual reasoning, and scene understanding. These abilities naturally align with the needs of autonomous driving, where complex reasoning about visual scenes and dynamic entities is crucial for safe decision-making. However, most general-purpose VLMs [40–43] are trained on web-scale data and are not specifically optimized for autonomous driving tasks. While they excel in generic perception and reasoning, they lack driving-specific priors and thus struggle to attend to critical visual cues and make safety-critical decisions [44]. Moreover, many existing VLMs are limited in their input structure: they either accept only a single image or treat multiple images as independent inputs without modeling their spatial or temporal relationships. This design restricts the model's ability to reason over surround-view inputs, which are essential for holistic scene understanding in autonomous driving. As a result, they struggle to integrate information across views, limiting cross-camera context awareness critical for safe decision-making.

To address these challenges, we adopt Senna-VLM [34], a recent VLM architecture specifically tailored for autonomous driving. Senna-VLM [34] introduces a Driving Vision Adapter to efficiently encode multi-view camera inputs and reduce token overload, and employs a multi-view prompt design to better contextualize spatial information for downstream reasoning. While Senna-VLM demonstrates strong performance under clean conditions, it still faces two key limitations: (1) limited robustness to corrupted visual inputs, and (2) insufficient diversity in QA formats, particularly multiple-choice questions (MCQs). To address these gaps, we further fine-tune Senna-VLM using two complementary QA datasets. DriveLM [28] enhances VLM capabilities for autonomous driving by introducing graph-structured

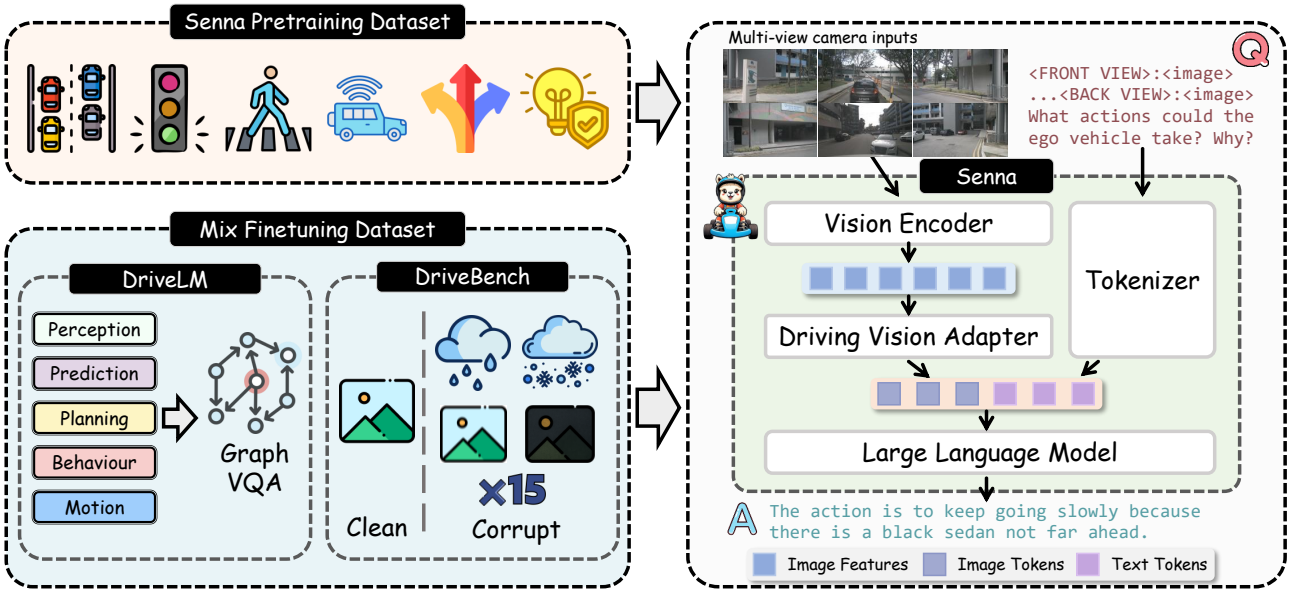

Figure 1. Illustration of our solution. We employ a driving-specific VLM, Senna-VLM, which takes multi-view images and a question as inputs and generates an answer. Senna-VLM is pretrained on the Senna dataset containing diverse safe-driving QA pairs. To further improve its capability on driving-related tasks and robustness against corrupted data, we fine-tune Senna-VLM using two additional datasets: DriveLM, which provides graph-structured VQA, and DriveBench, which introduces 15 types of scene corruptions.

reasoning and a corresponding Graph VQA dataset, which enables more coherent and context-aware decision-making. DriveBench [45] provides a large collection of multiple-choice questions together with 15 types of synthetic corruptions across diverse driving scenes. It also corrects the planning bias toward the "going ahead" action and filters out overly difficult cases, resulting in a cleaner and more balanced dataset for fine-tuning.

We validate our approach in the Track 1: Driving with Language of RoboSense Challenge 2025 [46], which benchmarks high-level reasoning in autonomous driving under both clean and corrupted multi-view camera inputs. Our method achieves a top-5 ranking in the final leaderboard, outperforming strong VLM baselines in multiple sub-tasks, including scene description, motion prediction, and planning under both clean and corruption cases. These results demonstrate the effectiveness of combining domain-specific architectural priors with diverse, corruption-aware QA supervision for building a robust and interpretable autonomous driving system.

2. Solution

The overview of our solution is depicted in Figure 1. To equip the autonomous driving system with stronger commonsense reasoning and a more comprehensive decision-making capability, we employ Senna-VLM, a LLaVA [40]-based vision-language model specifically designed for autonomous driving. To mitigate the impact of corrupted inputs on the VLM model, we leveraged a combination of publicly available driving QA datasets. These datasets, when aggregated, provide high-quality, diverse, and corruption-aware visual-question pairs that cover a broad spectrum of real-world driving scenarios. In the following subsections, we first introduce the Senna architecture in Sec. 2.1, then describe the training data we used in Sec. 2.2, and finally summarize our training strategy in Sec. 2.3.

2.1. Senna-VLM

Senna was originally developed to address the challenges of planning in autonomous driving, where the goal is to predict low-level driving trajectories. This task requires generating precise numerical outputs, a setting in which conventional vision-language models (VLMs) often fall short. To overcome this limitation, Senna adopts a cognitively inspired two-stage approach: rather than directly regressing trajectory coordinates, the Senna-VLM module first produces a high-level driving plan in natural language, such as commands like "turn left" or "accelerate", based on its perception of the scene. This intermediate representation is then mapped to low-level control signals through a lightweight regression head called Senna-E2E. Such a decomposition improves numerical accuracy and also enhances the interpretability of the model's decision-making process. Formally, the process

can be expressed as:

$$\hat{\tau} = \text{Senna-E2E}(\text{Senna-VLM}(I, Q)),$$

where $\hat{\tau} = \{(x_t, y_t)\}_{t=1}^{T}$ is the prediction low-level trajectories in T steps, I denotes the multi-view image inputs, Q is the task instruction or query. In this competition, we only focus on generating high-level decisions, we exclusively utilized the Senna-VLM module and omitted the end-to-end module.

Autonomous driving demands precise environmental understanding to support safe and reliable decision-making. VLMs through large-scale pretraining on image-text pairs, acquire rich commonsense knowledge that enables them to perceive complex road scenes, reason about the behavior of surrounding agents, and anticipate potential risks. Senna-VLM inherits these capabilities and is equipped for scene-level reasoning, including road layout interpretation, traffic participant recognition, and high-level plan formulation based on multimodal inputs. These strengths align closely with the requirements of this competition, making Senna-VLM a natural and effective choice for our system.

Senna-VLM follows the architecture of LLaVA-v1.6-vicuna 7B [40], consisting of a visual encoder E_{img}, a text tokenizer E_{txt}, and a large language model (LLM). However, a key distinction lies in its use of a Driving Vision Adapter (DVA), which not only projects visual features into the language feature space but is also specifically designed to enable more comprehensive perception of the surrounding environment.

While many existing open-source VLMs are limited to a single forward-facing image as input, this narrow field of view often fails to capture critical information from the sides or rear, potentially leading to unsafe or suboptimal decisions. To improve decision-making accuracy, it is essential for the model to incorporate multi-view visual inputs. Naively feeding all camera views into the model leads to an excessive number of image tokens, which not only increases training and inference latency but can also cause model collapse or decoding failure [34]. To address this, the Driving Vision Adapter compresses and projects the visual inputs into the LLM's feature space while reducing the number of image tokens. Specifically, each image is first transformed into a feature representation using a multi-layer perceptron composed of stacked linear layers followed by GELU activations [47]. A set of learnable image queries Q_{img} then attends to these features through a multi-head attention mechanism to produce a compact, semantically rich token representation.

The full processing pipeline of Senna-VLM is as follows. Given a set of surround-view images $\{I^{(v)}\}_{v=1}^{V} \in \mathbb{R}^{V\times H\times W\times 3}$, where H and W are the hight and width of each image. Each image is first resized to 224×224 and then passed through the visual encoder E_{img} to extract visual features: $F^{(v)} = E_{\text{img}}(I^{(v)})$. Senna-VLM adopts CLIP ViT-L/14 [48] as the backbone. The resulting feature maps $\{F^{(v)}\}$ are then processed by the Driving Vision Adapter (DVA), which performs both spatial compression and cross-view integration. Specifically, the DVA maps visual features into the language embedding space and outputs a compact sequence of image tokens:

$$T_{\text{img}}^{(v)} = \text{MHSA}(Q_{img}, W \cdot F^{(v)}, W \cdot F^{(v)}),$$

where W is the multi-layer perceptron. In parallel, the textual input Q (either a natural language instruction or a question) is tokenized using a text encoder E_{txt}, resulting in a sequence of text tokens: $T_{\text{txt}} = E_{\text{txt}}(Q)$. The final input to the language model is the concatenation of both modalities. This token sequence is fed into the LLM to generate the final output:

$$A = \text{LLM}(\text{Concat}(T_{\text{img}}, T_{\text{txt}})),$$

where A is the predicted high-level decision or answer to the input query. Senna-VLM utilizes Vicuna-v1.5-7B [49] as the language model.

To support effective multi-view reasoning, Senna-VLM also introduces a prompt formatting scheme that explicitly links each image to its respective viewpoint. Since questions in the task may refer to specific regions of the driving scene without explicitly naming the corresponding camera, the prompt helps the model align each image token with its context. The template is as follows: `<FRONT VIEW>:\n<image>\n`. Here, `<image>` is a special placeholder token within the LLM's vocabulary, which is replaced at runtime with the corresponding image token T_{img} after compression. This structured layout significantly enhances the model's ability to reason over multi-view input and respond to spatially grounded queries.

2.2. Dataset

In this section, we present all the datasets used in our pipeline, including those employed during both pretraining and fine-tuning stages.

2.2.1 Senna Dataset

To improve scene understanding and enable the acquisition of driving-related commonsense knowledge, Senna introduces a automatically generation pipeline to generate comprehensive question-answering (QA) dataset to train Senna-VLM. Following many planning-oriented frameworks, Senna decomposes the autonomous driving task into three key components: scene perception, motion prediction, and planning, which comprises six categories in total.

For the perception component, a primary focus is placed on understanding the scene context. To enable Senna-VLM to form a holistic understanding of the surrounding environment, the dataset includes a *Scene Description* category.

This encourages the model to answer high-level questions about key factors that may influence driving decisions, such as weather conditions, time of day, road surface status, and traffic density. In addition, traffic signals are crucial for safe navigation. The dataset therefore includes queries regarding the presence and state of traffic lights visible in the front-facing camera, helping VLM learn to avoid dangerous maneuvers when signals are present or ambiguous. Another essential factor is the presence of vulnerable road users (*e.g.*, pedestrians, cyclists). To ensure the model pays sufficient attention to such agents, dedicated QA pairs ask VLM to identify and enumerate all potentially vulnerable entities in the scene.

To support motion prediction, the dataset provides current location and velocity information of dynamic agents within the front-view frame. VLM is then required to predict their future trajectories and speed profiles in natural language, using simple semantic descriptors such as *"KEEP, STRAIGHT"*.

Finally, the planning component is the most critical in the overall architecture. Based on a complete understanding of the current environment, the model is tasked with generating a high-level plan for the ego vehicle, including future direction and speed intent, in the same semantic format as above. To ensure interpretability and reliability, the model must also provide a natural language justification for its decision.

Altogether, the QA pairs in the Senna dataset equip Senna-VLM with the ability to reason about the driving scene from perception to planning, enabling rational and interpretable decision-making under diverse traffic conditions.

Despite its comprehensive coverage of perception, prediction, and planning tasks, the original Senna dataset has certain limitations. All input images are clean and clear, making the trained model vulnerable to performance degradation under adverse weather conditions (*e.g.*, rain, fog) or low-quality visual inputs (*e.g.*, motion blur, lens obstacle). Additionally, the dataset exhibits limited linguistic diversity: each question follows a fixed template, and the answers are constrained to a narrow set of patterns. This lack of variety reduces the model's robustness to corrupted inputs and weakens its performance on tasks such as multiple-choice question answering (MCQ). To address these limitations, we incorporate two additional driving QA datasets to enhance both input robustness and linguistic diversity during fine-tuning.

2.2.2 Datasets for Robust and Diverse Fine-tuning

To address the limitations of the original Senna dataset we mentioned before, we further incorporate two comprehensive and high-quality driving QA datasets: DriveBench [45] and DriveLM [28].

To strengthen VLMs in autonomous driving tasks, DriveLM introduces the concept of graph-structured reasoning and constructs the corresponding Graph VQA dataset. Graph-structured reasoning enables step-by-step inference, where the answer to a previous question guides the reasoning for subsequent ones, ultimately leading to safe decision-making. In Graph VQA, all questions related to key objects within a single scene are organized into a directed acyclic graph, forming logical chains of reasoning. The dataset covers five core aspects of driving systems, arranged in a logical order: Perception, Prediction, Planning, Behavior, and Motion. Among them, Perception, Prediction, and Planning modules contain QA pairs associated with multiple key objects, while Behavior and Motion further refine the reasoning process to produce low-level trajectory outputs. The dataset provides a large and diverse set of questions, annotated with either human or model-generated ground truth, and carefully validated to ensure reliability, making it a robust benchmark for reasoning in autonomous driving.

DriveBench is constructed on top of DriveLM to evaluate the reliability of VLMs and improve their robustness under corrupted visual inputs. It inherits the same question type design as DriveLM while introducing an additional multiple-choice task for corruption type identification. To address the strong bias in DriveLM's planning MCQs, where "Going Straight" accounts for about 80%, DriveBench re-samples the data to achieve a more balanced distribution. Extremely challenging cases that require fine-grained visual cues or temporal context were removed to avoid unrealistic difficulty. To simulate diverse real-world degradations, DriveBench applies 15 corruption types generated with image processing algorithms, covering weather conditions, motion blur, sensor failures, transmission errors, and external disturbances. These augmentations provide realistic corrupted scenes and enhance the model's robustness against quality-degraded inputs.

By combining these two datasets, we construct a training set that is both visually and linguistically robust, thereby enhancing the capability of autonomous driving VLMs to make safer and more accurate decisions while remaining resilient to corrupted visual inputs.

2.3. Training Strategy

The model we used in experiment go through two stage of a structured pretraining phase followed by task-specific fine-tuning.

2.3.1 Pretraining Strategy

The pretraining process of Senna-VLM is organized into three stages, each designed to build specific capabilities in the model. To ensure that the DVA can effectively project image features into the language embedding space, the model is first trained on single-image QA data. This includes both

generic vision-language datasets and scene description QA pairs from the Senna dataset. To enhance the model's ability to reason over multiple camera views and perform image token compression, the second stage introduces surround-view inputs. This allows the DVA to learn better spatial integration. Finally, the model is further trained on meta-action planning examples from the Senna dataset, which require the generation of high-level plans based on complex driving scenes. This step strengthens the model's capacity to perform interpretable decision-making grounded in visual context.

2.3.2 Fine-tuning Strategy

For fine-tuning, we adopt a standard supervised fine-tuning (SFT) approach using a mixture of QA pairs from the DriveBench and DriveLM datasets. Given an image-text input pair (x, y), where $x = \text{Concat}(T_{\text{img}}, T_{\text{txt}})$ and y is the expected textual answer or planning decision, the training objective minimizes the negative log-likelihood:

$$\mathcal{L}_{\text{SFT}} = -\sum_{t=1}^{T} \log p(y_t \mid y_{<t}, x; \theta),$$

where θ denotes the parameters of the Senna-VLM, and y_t is the token at position t in the target sequence. To reduce computational and memory costs during training, we employ Low-Rank Adaptation (LoRA) [50] for parameter-efficient fine-tuning.

3. Experiments

3.1. Dataset

We use the official data provided by the *RoboSense Challenge 2025* [51] held at IROS 2025. This competition builds upon the legacy of the *RoboDepth Challenge 2023* [15, 52] at ICRA 2023 and the *RoboDrive Challenge 2024* [8, 53] 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 [45, 54–57]. Specifically, this task is built upon the **DriveBench** dataset [45] 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

To evaluate the effectiveness of our approach, we conduct experiments in two phases. Phase 1 focuses on Senna-VLM's ability to answer high-level driving questions from clean multi-view camera inputs. The questions cover perception, prediction, and planning tasks:

- **Perception-MCQs**
- **Perception-VQAs-Object-Description**
- **Perception-VQAs-Scene-Description**
- **Prediction-MCQs**
- **Planning-VQAs-Scene-Description**
- **Planning-VQAs-Object-Description**

Phase 2 evaluates Senna-VLM's robustness to corrupted inputs. We test whether the model can still answer the above high-level driving questions under corruption, and also assess its ability to identify corruption types through **Corruption-MCQs**. For both phases, we follow the official RoboSense Challenge 2025 evaluation protocol, combining MCQ accuracy and an LLM-based scoring metric for open-ended responses. To explore the impact of different data combinations and training strategies, we fine-tune models with multiple dataset settings and compare against a Qwen2.5-VL 7B [43] baseline (Qwen for short). All dataset combination and fine-tuning strategies are listed below.

Phase 1 strategies:

- **S1**: Senna pretrained model.
- **S2**: Fine-tune on QA pairs generated by the Senna pipeline based on the nuScenes [58] dataset for 1 epoch.
- **S3**: Fine-tune on DriveBench (clean inputs only) for 1 epoch.
- **S4**: Fine-tune on DriveBench (clean inputs only) for 2 epochs.
- **S5**: Fine-tune on DriveBench (clean inputs only) for 5 epochs.
- **S6**: Fine-tune on one-sixth of DriveLM QA pairs combined with duplicated DriveBench with only clean inputs for 1 epoch.

Phase 2 strategies:

- **S6**: Use the best performaning model in phase 1 and test its performance under corruption.
- **S7**: Fine-tune on DriveBench (clean + 15 corruptions) for 2 epochs.
- **S8**: Fine-tune on DriveBench (clean + 15 corruptions) plus the entire DriveLM QA pairs for 1 epoch.

3.3. Implementation Details

We initialize Senna-VLM from the official Senna pretrained weights, which were trained on the large-scale DriveX dataset [34]. We then fine-tune the model on mixtures of DriveLM [28] and DriveBench [45] according to the above strategies. All fine-tuning hyperparameters follow the official Senna settings: LoRA rank of 128, $\alpha = 256$, learning rate of 2×10^{-5}, batch size of 8, and a cosine learning-rate scheduler. All experiments are conducted on a single NVIDIA H100 GPU. Evaluation is performed on the official RoboSense Challenge test set, which consists of QA pairs derived from nuScenes.

Table 1. Results of phase 1 show the performance of different fine-tuning strategies under clean input settings across multiple driving-related question-answering tasks. Here, the baseline refers to the zero-shot performance of Qwen2.5-VL 7B [43].

Question Type	Baseline	S1	S2	S3	S4	S5	S6
Perception-MCQs	75.47	20.75	13.21	79.25	83.02	77.36	71.70
Perception-VQAs-Object-Description	29.02	16.93	18.57	27.48	33.34	31.23	42.93
Perception-VQAs-Scene-Description	22.50	35.23	33.20	21.25	22.97	35.00	34.06
Prediction-MCQs	59.20	56.90	59.00	59.20	59.20	59.20	59.20
Planning-VQAs-Scene-Description	29.85	36.11	39.91	41.53	43.02	40.75	53.74
Planning-VQAs-Object-Description	31.24	26.36	28.63	52.73	51.00	51.91	58.24
Score	42.50	37.49	39.20	48.96	49.95	49.87	54.79

Table 2. Results of phase 2 present the evaluation of model robustness to corrupted visual inputs across multiple driving-related question-answering tasks, covering both high-level driving questions and corruption-aware MCQs. Here, the baseline refers to the zero-shot performance of Qwen2.5-VL 7B [43].

Question Type	Baseline	S6	S7	S8
Perception-MCQs	78.57	0.00	37.76	70.41
Perception-VQAs-Object-Description	21.74	26.40	31.23	41.66
Perception-VQAs-Scene-Description	19.31	41.99	43.73	37.15
Prediction-MCQs	61.56	61.56	61.56	65.51
Corruption-MCQs	30.83	58.65	89.42	50.00
Planning-VQAs-Scene-Description	31.25	49.19	42.40	51.38
Planning-VQAs-Object-Description	81.73	49.18	45.05	48.34
Score	43.72	47.70	49.72	53.98

3.4. Experiment Results

Tables 1 and 2 summarize the results for Phase 1 and Phase 2, respectively. In Phase 1, both the model using Senna's pre-trained weights (S1) and the model fine-tuned on Senna QA pairs (S2) underperform on Perception-MCQs, likely due to the lack of multiple-choice supervision in Senna's QA generation pipeline, resulting in performance even below the Qwen baseline. However, since the Senna data generation pipeline includes a rich set of scene description questions, both S1 and S2 perform better on Perception-VQAs-Scene-Description and Planning-VQAs-Scene-Description, and achieve comparable results on Prediction-MCQs. Nonetheless, their performance on other tasks remains limited, suggesting that the limited diversity of questions and answers in Senna may hinder its generalization to broader task categories. When fine-tuned on DriveBench with only clean inputs only (S3, S4, S5), MCQ accuracy improves markedly with different epoch settings, reaching 77.36, 79.25, and 83.02, which clearly surpasses both Senna-trained models and Qwen. Performance on Perception-VQAs-Scene-Description is also comparable or better, despite the much smaller training volume (only tens of minutes to a few hours). We find that S4 achieves the best trade-off between training time and performance. Incorporating a subset of DriveLM (S6) further improves performance on most tasks, particularly in the two Planning categories, with gains of approximately 10 and 6 points over the best DriveBench-only results. This yields a higher overall average score of 54.79, surpassing the baseline by 12.3 points.

In Phase 2, we first evaluate the S6 model under corruption. Although it achieved 71.7 on Perception-MCQs with clean inputs, it drops to 0 under corrupted inputs, indicating a complete lack of robustness. Its Corruption-MCQ scores also lag behind Qwen. Nevertheless, in other tasks, the model still outperforms the zero-shot baseline. We then train on DriveBench including all 16 input types (S7). This leads to a substantial improvement on Perception-MCQs, increasing from 0 to 37.76, though it remains substantially lower than Qwen's score of 78.57. Crucially, this model attains the highest Corruption-MCQ score of 89.42, showing that exposure to diverse corruptions greatly improves discrimination of corruption types. Finally, combining all QA pairs from DriveLM with the full DriveBench dataset (S8) results in the widest coverage and highest diversity among all training configurations. This model effectively balances corruption robustness with enriched driving knowledge, achieving strong performance on Perception-MCQs and competitive scores across other tasks. However, it still does not surpass Qwen in corruption robustness, likely because Qwen's large-scale pretraining corpus includes a greater number

of corrupted visual samples and more extensive MCQ supervision. Unexpectedly, this combined model obtains the lowest Corruption-MCQ score, possibly because the larger volume of QA pairs from DriveLM dilutes the corruption-specific signal provided by DriveBench, thereby weakening the model's ability to discriminate between corruption types. Nevertheless, owing to its high Perception-MCQ score and consistently strong performance across other categories, this model achieves the highest overall average score among all our submissions.

4. Conclusion

To enable an autonomous driving system with stronger commonsense reasoning and improved instruction-following capabilities, we adopted Senna-VLM, a high-performing vision-language model specifically designed for driving scenarios. In particular, the Driving Vision Adapter and multi-view prompt design in Senna-VLM allow the model to better perceive and reason about the ego vehicle's surrounding environment, leading to more reliable and interpretable planning decisions. To further enhance the model's robustness to corrupted inputs and improve its accuracy on multiple-choice question answering, we fine-tuned Senna-VLM using a combination of high-quality driving QA datasets, namely DriveLM and DriveBench. Our solution achieved a top-5 performance in Track 1 of the RoboSense Challenge, demonstrating the effectiveness of our approach.

References

[1] Li Chen, Penghao Wu, Kashyap Chitta, Bernhard Jaeger, Andreas Geiger, and Hongyang Li. End-to-end autonomous driving: Challenges and frontiers. *IEEE Trans. Pattern Anal. Mach. Intell.*, 46(12):10164–10183, 2024.

[2] Jinkyu Kim, Teruhisa Misu, Yi-Ting Chen, Ashish Tawari, and John F. Canny. Grounding human-to-vehicle advice for self-driving vehicles. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 10591–10599. Computer Vision Foundation / IEEE, 2019.

[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] Jiageng Mao, Yuxi Qian, Hang Zhao, and Yue Wang. Gpt-driver: Learning to drive with GPT. *CoRR*, abs/2310.01415, 2023.

[5] 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.

[6] 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.

[7] Jyh-Jing Hwang, Runsheng Xu, Hubert Lin, Wei-Chih Hung, Jingwei Ji, Kristy Choi, Di Huang, Tong He, Paul Covington, Benjamin Sapp, Yin Zhou, James Guo, Dragomir Anguelov, and Mingxing Tan. EMMA: end-to-end multimodal model for autonomous driving. *Trans. Mach. Learn. Res.*, 2025, 2025.

[8] 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.

[9] 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.

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

[12] 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.

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

[14] 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.

[15] 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.

[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, 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.

[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] 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.

[20] 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.

[21] Chenyu Yi, Siyuan Yang, Haoliang Li, Yap-Peng Tan, and Alex C. Kot. Benchmarking the robustness of spatial-temporal models against corruptions. In Joaquin Vanschoren and Sai-Kit Yeung, editors, *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual*, 2021.

[22] 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.

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

[24] 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.

[25] 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.

[26] 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.

[27] 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.

[28] 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 Ales Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol, editors, *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part LII*, volume 15110 of *Lecture Notes in Computer Science*, pages 256–274. Springer, 2024.

[29] Xiaoyu Tian, Junru Gu, Bailin Li, Yicheng Liu, Chenxu Hu, Yang Wang, Kun Zhan, Peng Jia, Xianpeng Lang, and Hang Zhao. Drivevlm: The convergence of autonomous driving and large vision-language models. *CoRR*, abs/2402.12289, 2024.

[30] 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.

[31] 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.

[32] 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.

[33] Chenbin Pan, Burhaneddin Yaman, Tommaso Nesti, Abhirup Mallik, Alessandro Gabriele Allievi, Senem Velipasalar, and Liu Ren. VLP: vision language planning for autonomous driving. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 14760–14769. IEEE, 2024.

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

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

[36] 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.

[37] 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.

[38] 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.

[39] Bo Jiang, Shaoyu Chen, Qian Zhang, Wenyu Liu, and Xinggang Wang. Alphadrive: Unleashing the power of vlms in autonomous driving via reinforcement learning and reasoning. *CoRR*, abs/2503.07608, 2025.

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

[41] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, January 2024.

[42] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. *CoRR*, abs/2312.14238, 2023.

[43] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 2023.

[44] Xiaosong Jia, Zhenjie Yang, Qifeng Li, Zhiyuan Zhang, and Junchi Yan. Bench2drive: Towards multi-ability benchmarking of closed-loop end-to-end autonomous driving. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang, editors, *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024.

[45] 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.

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

[47] Dan Hendrycks and Kevin Gimpel. Bridging nonlinearities and stochastic regularizers with gaussian error linear units. *CoRR*, abs/1606.08415, 2016.

[48] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR, 2021.

[49] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine, editors, *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023.

[50] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022.

[51] 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.

[52] 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.

[53] 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.

[54] 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.

[55] 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.

[56] 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.

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

[58] 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. *arXiv preprint arXiv:1903.11027*, 2019.