Toward Automatic Safe Driving Instruction: A Large-Scale Vision Language Model Approach

Haruki Sakajo[1], Hiroshi Takato[2] Hiroshi Tsutsui[2] Komei Soda[2,3]
Hidetaka Kamigaito[1], Taro Watanabe[1]
[1]Nara Institute of Science and Technology (NAIST),
[2]Teatis inc., [3]Queensland university of technology
`sakajo.haruki.sd9@naist.ac.jp`
`{tak, james}@dotsfty.com` `komei.soda@connect.qut.edu.au`
`{kamigaito.h, taro}@is.naist.jp`

Abstract

Large-scale Vision Language Models (LVLMs) exhibit advanced capabilities in tasks that require visual information, including object detection. These capabilities have promising applications in various industrial domains, such as autonomous driving. For example, LVLMs can generate safety-oriented descriptions of videos captured by road-facing cameras. However, ensuring comprehensive safety requires additional monitoring driver-facing views to detect risky events, such as the use of mobiles while driving. Thus, the ability to process synchronized inputs is necessary from both driver-facing and road-facing cameras. In this study, we develop a model integrating two video inputs and investigate the capabilities of LVLMs by constructing a dataset and evaluating their performance on this dataset. Our experimental results demonstrate that while pre-trained LVLMs have limited effectiveness, fine-tuned LVLMs can generate accurate and safety-aware driving instructions. Nonetheless, several challenges remain, particularly in detecting subtle or complex events in the video. Our findings and error analysis provide valuable insights that can contribute to the improvement of LVLM-based systems in this domain.

1 Introduction

The promising capabilities of Large Language Models (LLMs) are changing this society by assisting various tasks, e.g., coding (Rozière et al., 2024) and education (Liu et al., 2024). Large-scale Vision Language Models (LVLMs) possess high capabilities in the intersection of vision and language tasks, leveraging the capabilities of LLMs, such as inference and instruction following, by integrating a vision encoder. Therefore, LVLMs have been adopted across domains that require both visual and textual information, including the medical application (Li et al., 2023a; Yan et al., 2024; Pal and Sankarasubbu, 2024) and driving assistance (Arai et al., 2025; Duan et al., 2024; Xuan et al., 2024).

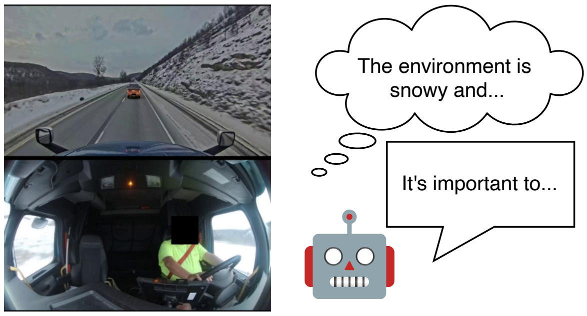

Figure 1: Illustration of an application of this study. A Model provides driving instructions for the given video.

In a driving domain, LVLMs are used to detect objects on the road, generate actions, and provide safe driving instructions (Zhou et al., 2024; Lu et al., 2025; Duan et al., 2024; Xuan et al., 2024). With the rapid growth of the dashcam industry, LVLMs are increasingly exploited to interpret driving scenes captured by the dashcam. Previous studies investigated the capabilities of LVLMs to interpret vehicle behavior and suggest actions for safe driving (Duan et al., 2024; Xuan et al., 2024). However, while a single dashcam for the road-facing view is enough to detect risky actions of vehicles (e.g., harsh turns), a driver-facing view using an additional dashcam is needed, for example, when detecting the driver's use of mobiles while driving.

In this study, we develop LVLM-based models that generate instructions for safe driving based on two video streams that synchronize driver-facing and road-facing views. We collect such videos and annotate each clip in two stages: first, by detecting events that happened in them, and second, by generating safe driving instructions. We create a conversation-style dataset, where a user asks what is happening in the video, followed by a request to generate safe driving instructions. Our experiments demonstrate that fine-tuned LVLMs on our dataset can generate safe driving instructions that are well-aligned with the visual context, whereas pre-trained

Proceedings of MMLoSo 2025: First Workshop on Multimodal Models for Low-Resource Contexts and Social Impact, pages 13–24
December 23, 2025

Figure 2: Overview of our dataset construction approach. The dataset contains videos with synchronized driver-facing and road-facing views. GPT-4o generates the gold answers to event detection and safe driving instruction questions based on annotated labels for each video.

Human	What is happening in the video?
Assistant	The environment is icy with clear weather as you approach an intersection. A rolling stop occurs as you navigate the intersection at 8 mph while properly using your turn signal.
Human	What is recommended actions for the ego-car?
Assistant	\ud83d\udca1Paying extra attention to road conditions due to ice is important. Make sure to bring your vehicle to a complete stop at intersections to ensure safety, especially in icy conditions where stopping distances are longer.

Table 1: An example of the question and its answer. The top is the event detection, and the bottom is the safe driving instruction.

models generate generic guidance. Nonetheless, even the fine-tuned models occasionally fail to detect risky events (e.g., harsh turns). Our findings and analysis provide valuable insights that help improve LVLM-based driving instruction systems, as illustrated in Figure 1.

2 Background and Related Work

2.1 Large-scale Vision Language Models

Large-scale Vision Language Models (LVLMs), such as Flamingo (Alayrac et al., 2022), BLIP-2 (Li et al., 2023b), and LLaVA (Liu et al., 2023), integrate a vision encoder with a Large Language Model (LLM), enabling them to process visual inputs (e.g., images) and achieving advanced performance on Visual Question–Answering (VQA) tasks (Liu et al., 2023; Bai et al., 2025). Visual instruction tuning is also effective for further performance improvement (Liu et al., 2023). The enhanced capabilities of LVLMs are helpful across various domains, including disease detection from clinical images (Yan et al., 2024; Pal and Sankarasubbu, 2024), artwork explanation (Hayashi et al., 2024; Ozaki et al., 2025), and vowel prediction from MRI-based articulatory images and videos (Sakajo et al., 2025).

2.2 Language Models in Driving Scenario

LLMs and LVLMs are also helpful in driving domains, and several studies have demonstrated their capabilities (Zhou et al., 2024). For example, LVLMs visually understand traffic signs (Lu et al., 2025), and LLMs can be utilized to develop models for processing LiDAR data (Yang et al., 2023) and autonomous driving (Xu et al., 2024). LVLMs are also leveraged for constructing a driving dataset (Arai et al., 2025). In the AI City Challenge (Wang et al., 2024), LVLM-based approaches (Duan et al., 2024; Xuan et al., 2024) achieved advanced performance for the traffic safety description and analysis task, suggesting that LVLMs have the capabilities to learn and provide descriptions regarding safe driving. However, the capabilities of LLMs to instruct safe driving with synchronized driver-facing and road-facing RGB cameras are unexplored.

3 Dataset Construction

We constructed a dataset to evaluate LVLMs in the context of safe driving instructions, as illustrated in Figure 2. The dataset comprises videos and questions, along with expert-annotated answers.

Primary Event	Description
Crash	Any visible crash involving ego-vehicle or other vehicles.
Forward Collision Warning	An event where the ego-vehicle is at risk of imminent front-end collision.
Tailgating (Following Distance)	Following another vehicle at a dangerously short distance.
Harsh Brake	A sudden, strong deceleration by the ego-vehicle.
Harsh Turn	A sharp, abrupt turn indicating potential loss of control or evasive maneuver.
Rolling Stop	Failure to come to a complete stop at stop signs or similar control points.
Mobile Usage	The driver is observed using a mobile phone (hands-on or hands-free).
Inattentive Driving	Observable distraction or loss of attention by the driver.
Lane Departure	Vehicle crosses out of its lane without clear intention or necessity.
Other Events	Includes seatbelt violations or distraction events.

Table 2: Primary Event Categories for Annotation.

Sub-Event	Options
Lane Cut Off	Proper use of turn signal,Improper use of turn signal
Lane Change	Proper use of turn signal, Improper use of turn signal, To avoid primary event, Root cause of primary event
Turn (Other Vehicles)	Proper use of turn signal, Improper use of turn signal
Turn (Ego Vehicle)	Proper use of turn signal (listen to the audio), Improper use of turn signal
Signs of Aggressive Reaction	Vehicle maneuver, Aggressive language, Honk, None, Unknown
Signs of Distraction	Smoking, Mobile phone, Playing with hair, Drinking, Eating, Picking something from the floor, Reaching behind the backseat, Yawning, None, Unknown
Weather Condition	Clear, Rainy, Foggy, Snowy
Road Condition	Dry, Wet, Icy
Visibility Condition	Clear, Poor
Road Information	Highway, Highway merge, Local road, Intersection, 3-leg intersection, School zone, Construction zone, Residential area, Rural roads, Tunnel, Pedestrian crossroad
Speed Management	Decrease, Maintain, Increase

Table 3: Sub-Event Categories and Options.

	Train	Validation	Test
Samples	1,719	215	215
Duration (s)	18,720	2,311	2,371
Frames	561,223	69,291	72,836

Table 4: Dataset statistics.

3.1 Task

We evaluate LVLMs' capabilities and challenges using a conversation-style VQA task related to safe driving. We provide LVLMs with synchronized driver-facing and road-facing videos captured using RGB cameras and then ask the LVLMs to explain what happens in the video and generate instructions for safe driving.

3.2 Video Collection

We collected vehicle speed and video recordings from both driver-facing and road-facing RGB cameras and lined them up vertically as unified clips, placing the road-facing view on the top and the driver-facing view on the bottom. Each video in our dataset presents synchronized views of both the driver and the road.

Parameter	Value
Batch size	8
Epoch	3
Learning rate	1e-5
Learning rate scheduler	cosine
Adam β_1	0.9
Adam β_2	0.999
Adam ϵ	1e-8
Precision	BF16
Video Max Pixels	16,384
Video Min Pixels	256
Video Maxlen	128
Video FPS	2
Seed	42

Table 5: Hyperparameters for fine-tuning.

3.3 Question Definition

We adopt a Chain-of-Thought (Kojima et al., 2022) and a conversation-style template to facilitate step-by-step reasoning. The dataset has two questions: (1) *"What is happening in the video?"* and (2) *"What is recommended actions for the ego-car?"*. We refer to the first type as **event detection** questions/answers and the second as **safe driving instruction** questions/answers. An example of each type of question–answer pair is presented in Table 1.

This Chain-of-Thought format is designed to guide the model through a reasoning process that first identifies events in the video and then infers appropriate driving actions based on those observations.

3.4 Annotations

To support structured, context-rich labeling of driving scenarios, we implemented a three-step process.

Primary event selection. An annotator begins by selecting a single **primary event** from a predefined taxonomy of safety-critical driving behaviors, as listed in Table 2. These events capture the core nature of the incident.

Sub-event selection. Next, an annotator is encouraged to select as many relevant **sub-events** as necessary to describe the contributing context. These sub-events, summarized in Table 3, include surrounding vehicle behaviors (e.g., lane changes, turn maneuvers), environmental conditions (e.g., weather, visibility), or behavioral cues (e.g., signs of aggression). This multi-label scheme enables fine-grained characterization of complex traffic scenes.

Summary generation. In the final step, a natural language annotation is automatically generated using GPT-4o (OpenAI et al., 2024). The model takes as input the selected primary and sub-events, along with auxiliary data such as the vehicle's speed at the time of the event. Based on this information, GPT-4o generates a descriptive summary that answers two key questions: "What is happening in the video?" and "What is recommended actions for the ego-car?" Finally, experts manually review the generated descriptions and confirm the quality.

3.5 Data Statistics

Table 4 shows our dataset statistics. A video has an approximate duration of 10 seconds and 30 frames per second. The number of primary events and sub-event options is provided in Appendix A.

4 Experimental Settings

4.1 Dataset

We use our dataset introduced in Section 3 and treat the videos as two frames per second.

4.2 Models

In this study, we utilize Qwen2.5-VL (Bai et al., 2025) 3B and 7B models, which are available even with limited computational resources. We also fine-tune these models on our dataset. We refer to fine-tuned Qwen2.5-VL-3B and fine-tuned Qwen2.5-VL-7B as Qwen2.5-VL-3B (FT) and Qwen2.5-VL-7B (FT), respectively.

4.3 Training and Inference

Models receive instruction and video inputs, while auxiliary sensor data, e.g., vehicle speed, was incorporated during dataset construction. This approach reflects the practical consideration that dashcams are easily deployable, whereas sensor installation requires additional costs.

Training. We freeze the vision encoder and train only the language model with full-parameter supervised fine-tuning. LVLMs fine-tuning is conducted using the LLaMA-Factory (Zheng et al., 2024) with DeepSpeed ZeRO stage 2 (Rajbhandari et al., 2020). We fine-tuned LVLMs using eight NVIDIA A100-SXM4-40GB GPUs and used LLaMA-Factory version 0.9.2.dev0 with minor modifications to load models correctly. Table 5 provides the hyperparameters.

Inference. We test LVLMs and fine-tuned LVLMs under the zero-shot setting. The evaluations are performed on an NVIDIA L4 GPU.

4.4 Metrics

We evaluate the quality of generated text by comparing it to the reference text in the dataset using BERTScore (Zhang* et al., 2020) and BLEU socres (Papineni et al., 2002) as evaluation metrics. We use the original implementation[1] for BERTScore using RoBERTa (Liu et al., 2019) and sacreBLEU (Post, 2018)[2] for BLEU scores.

5 Result and Discussion

Table 6 shows the results of each model on our dataset, and Tables 7 and 8 show the samples of generated text for event detection and safe driving instruction. Before fine-tuning, Qwen2.5-VL-3B performs better in terms of F1 score on BERTScore for the safe driving instruction than Qwen2.5-VL-7B, while the 7B model outperforms the 3B model in event detection. This suggests that the parameter size is irrelevant to the task performance of pre-trained models. Fine-tuning improves both BERT scores and BLEU scores, indicating that models can learn this task correctly.

[1] https://github.com/Tiiiger/bert_score
[2] https://github.com/mjpost/sacrebleu

Model	Event Detection				Safe Driving Instruction			
	P	R	F1	BLEU	P	R	F1	BLEU
Qwen2.5-VL-3B*†	0.8180*†	0.8504*†	0.8337*†	1.44*†	0.8273*†	0.8590*†	0.8426*†	1.66*†
Qwen2.5-VL-7B	0.8235*†	0.8498*†	0.8364*†	1.26*†	0.8086*†	0.8662*†	0.8364*†	1.15*†
Qwen2.5-VL-3B (FT)	0.8901*†	0.9041*†	0.8970*†	10.59*	0.8935*†	0.8954*†	0.8944*†	11.83*†
Qwen2.5-VL-7B (FT)	0.8999*†	0.9070*†	0.8983*†	10.24*	0.8994*†	0.9002*†	0.8997*†	13.40*†

Table 6: Results of event detection and safe driving instruction tasks. Metrics include BERTScore Precision (P), Recall (R), F1, and BLEU. We performed statistical significance tests using 10,000 bootstrap resamples with a significance level of $p = 0.05$. * and † indicate statistically significant differences between fine-tuned and non-fine-tuned conditions and across models, respectively. The performance differences across models and between fine-tuning and non-fine-tuning conditions are statistically significant, except for the BLEU scores for the event detection between Qwen2.5-VL-3B (FT) and Qwen2.5-VL-7B (FT)

Question	What is happening in the video?
Ground-truth	The environment is icy with clear weather as you approach an intersection. A rolling stop occurs as you navigate the intersection at 8 mph while properly using your turn signal.
Qwen2.5-VL 3B	The video shows a split-screen view of a snowy street scene and the interior of a vehicle. The top half of the screen captures the outside view, showing a snow-covered road with tire tracks, buildings on either side, and a stop sign ahead. The bottom half of the screen shows the inside of the vehicle, where a person is driving. The driver appears to be looking straight ahead, and there is a piece of paper or document on the dashboard. The scene suggests that the driver is navigating through a snowy environment, possibly preparing to make a turn or stop at the stop sign. [BERTScore] P: 0.8202, R: 0.8554, F: 0.8374 [BLEU] 0.5540
Qwen2.5-VL 7B	The video shows a first-person perspective from inside a vehicle, likely a truck or bus, as it drives through a snowy town. The view alternates between the road ahead and the interior of the vehicle, where a person is driving. The road appears to be partially cleared of snow, with patches of snow on the sides. There are buildings on both sides of the road, some of which have signs indicating they are businesses. The driver is focused on the road, and there are other vehicles visible in the background, including a white car and an orange vehicle parked near a building labeled "Body Shop." The weather seems overcast, and the ground is covered with snow. [BERTScore] P: 0.8139, R: 0.8439, F: 0.8286 [BLEU] 0.4693
Qwen3.5-VL 3B (FT)	The weather is clear and the road conditions are dry as you approach an intersection. You perform a rolling stop at the intersection, maintaining a speed of 10 mph. There are no signs of distraction or aggressive reaction from your side. [BERTScore] P: 0.8997, R: 0.9149, F: 0.9072 [BLEU] 14.3398
Qwen2.5-VL 7B (FT)	The environment around you is snowy with icy road conditions, making it challenging to drive safely. You approach an intersection and perform a rolling stop at 10 mph while using aggressive language. [BERTScore] P: 0.9132, R: 0.9202, F: 0.9167 [BLEU] 7.6260

Table 7: Sample ground-truth and generated sentences of Qwen2.5-VL 3B in event detection with evaluation metrics include BERTScore Precision (P), Recall (R), F1, and BLEU.

5.1 Result Comparison

Parameter size influence. The difference in the F1 score of BERTScore between Qwen2.5-VL-3B and Qwen2.5-VL-7B on the event detection was not significant, indicating that the performance of the event detection task is saturated in the zero-shot setting. The results demonstrate that the 3B model performs better for safe driving instruction than the 7B model, which can be attributed to the output

Question	What is recommended actions for the ego-car?
Ground-truth	\ud83d\udca1Paying extra attention to road conditions due to ice is important. Make sure to bring your vehicle to a complete stop at intersections to ensure safety, especially in icy conditions where stopping distances are longer.
Qwen2.5-VL 3B	Based on the video description, the recommended action for the ego-car (the vehicle being driven) would be to follow the instructions given by the driver. Since the driver is looking straight ahead and appears to be focused on the road, it is likely that they have made a decision about the next move. The driver may be preparing to make a turn or stop at the stop sign. Therefore, the ego-car should follow the driver's instructions and navigate accordingly. This could involve accelerating, decelerating, or changing lanes as directed by the driver. [BERTScore] P: 0.8356, R: 0.8400, F: 0.8378 [BLEU] 1.0798
Qwen2.5-VL 7B	In the context of the video, the ego-car (the vehicle being driven) is navigating a snowy environment. Here are some recommended actions that could help ensure safe driving: 1. **Drive Slowly**: Snowy roads can be slippery, so reducing speed helps maintain better control of the vehicle. 2. **Maintain a Safe Following Distance**: Increase the distance between your vehicle and the one in front to allow more time to react to any sudden stops or changes in traffic conditions. 3. **Use Low Beams**: If visibility is reduced due to snow or ice, use low beams to improve visibility without causing glare for other drivers. 4. **Avoid\u6025\u8f6c\u5f2f (Sharp Turns)**: Snow can cause the tires to lose traction, making sharp turns particularly dangerous. Try to avoid them if possible. 5. **Check Tires**: Ensure that your tires have adequate tread depth and are properly inflated. Snow chains may also be necessary in very snowy conditions. 6. **Be Aware of Other Vehicles**: Watch out for other vehicles, especially those that might be struggling to navigate the snow. Be prepared to stop or slow down if necessary. 7. **Stay Alert**: Snowy conditions can be deceptive, and visibility might be reduced. Stay alert and be ready to react to any changes in the road or traffic. 8. **Use Turn Signals**: Signal your intentions clearly to other drivers, especially when changing lanes or turning. 9. **Keep a Full Tank of Fuel**: Snowy conditions can lead to unexpected delays, so having enough fuel is crucial. 10. **Plan Your Route**: If possible, plan your route in advance to avoid areas with heavy snow accumulation or icy patches. By following these recommendations, the ego-car can navigate the snowy conditions safely and efficiently. [BERTScore] P: 0.8089, R: 0.8662, F: 0.8366 [BLEU] 0.4625
Qwen2.5-VL 3B (FT)	While it's good that you're paying attention to the road, it's important to come to a complete stop at intersections to ensure safety and compliance with traffic laws. Practice making full stops to improve your driving skills and enhance safety. [BERTScore] P: 0.9062, R: 0.8896, F: 0.8978 [BLEU] 22.1907
Qwen2.5-VL 7B (FT)	\ud83d\udca1In icy and snowy conditions, it's crucial to come to a complete stop at intersections to ensure safety. Try to remain calm and avoid aggressive language as it can distract from the primary task of driving safely. [BERTScore] P: 0.9142, R: 0.9243, F: 0.9192 [BLEU] 22.2504

Table 8: Sample ground-truth and generated sentences of Qwen2.5-VL 3 in safe drivinginstruction with evaluation metrics include BERTScore Precision (P), Recall (R), F1, and BLEU.

tendencies in the 7B model. The 7B model outputs general recommendations for safe driving instruction before fine-tuning, as shown in Table 9, resulting in lower precision and higher recall. For further analysis, we also computed self-BLEU scores (Zhu et al., 2018) for each event using the outputs generated by each model to assess diversity. The self-BLEU scores, as shown in Table 10, also indicate

Model & Event Type	Top-10 4-gram words
Qwen2.5-VL-3B	
Event Detection	half of the video; The video shows a; top half of the; The video shows two; the interior of the; The top half of; bottom half of the; The bottom half of; The interior of the; interior of the vehicle.
Safe Driving Instuction	Based on the video; on the video description,; the ego-car (the vehicle; for the ego-car (the; there are no specific; are no specific actions; recommended actions for the; the video description, there; video description, there are; a safe distance from.
Qwen2.5-VL-3B (FT)	
Event Detection	no signs of distraction; signs of distraction or; of distraction or aggressive; There are no signs; are no signs of; The footage shows you; footage shows you driving; distraction or aggressive reaction; or aggressive reaction from; aggressive reaction from your.
Safe Driving Instruction	While it's good that; come to a complete; to a complete stop; a complete stop at; it's good that you're; to come to a; it's important to come; important to come to; complete stop at intersections; stop at intersections to.
Qwen2.5-VL-7B	
Event Detection	The video appears to; half of the screen; of the screen shows; the screen shows the; video appears to be; the interior of the; appears to be a; shows the interior of; to be a split-screen; be a split-screen view.
Safe Driving Instruction	In the context of; the context of the; are some general recommendations; some general recommendations for; the ego-car (the vehicle; context of the video,; a safe distance from; ego-car (the vehicle being; distance from the vehicle; for the ego-car (the.
Qwen2.5-VL-7B (FT)	
Event Detection	signs of distraction or; of distraction or aggressive; no signs of distraction; There are no signs; are no signs of; clear weather and dry; weather and dry road; The footage shows you; footage shows you driving; and dry road conditions.
Safe Driving Instruction	come to a complete; to a complete stop; a complete stop at; increase your following distance; maintain a safe following; your following distance to; to maintain a safe; a safe following distance; important to maintain a; safe following distance to.

Table 9: The top 10 4-grams in each response.

	Detection	Instruction
Qwen2.5-VL-3B	84.4359	82.8134
Qwen2.5-VL-7B	83.6289	85.7856
Qwen2.5-VL-3B (FT)	96.4625	95.8846
Qwen2.5-VL-7B (FT)	95.5442	93.5108

Table 10: Self-BLEU scores. "Detection" and "Instruction" denote "Event Detection" and "Safe Driving Instruction", respectively.

that the 7B model outputs less diverse texts for the safe driving instruction when compared with the 3B model.

Performance improvement by fine-tuning. Fine-tuning improves overall performance, and Qwen2.5-VL-7B (FT) outperforms Qwen2.5-VL-3B (FT) on both tasks in terms of BERTScore, while Qwen2.5-VL-3B outperforms Qwen2.5-VL-7B before fine-tuning. Figures 3, 4 5 and 6 also show that fine-tuning improves overall performance. On the other hand, Table 6 shows that the difference in the BLEU scores between both fine-tuned models on the event detection task is not significant. This suggests that a larger parameter size has a positive effect on the fine-tuning performance of LVLMs for this task. In contrast, the final performance after fine-tuning remains consistent across model sizes with respect to BLEU scores.

5.2 Error Analysis

We focus on the subset of samples for which BERTScore of the safe driving instruction falls within the bottom 25%. Approximately 4% of all the samples are shared across the bottom 25% subsets for all models, which we refer to as the "difficult subset". Within this subset, 33% of the samples are annotated as good driving, and another 33% involve scenarios where the ego-car is turning right and left. Although all models generated recommendations to improve already good driving behaviors, the suggestions for safer driving varied slightly, resulting in relatively low scores.

For the turning right and left scenarios, the gold answers typically recommend turning while reducing speed. However, even fine-tuned models produced irrelevant suggestions, such as mentioning a stop sign not presented in the video. These observations suggest that while fine-tuned models are capable of generating various safety-related suggestions, they still struggle to detect issues such as excessive speed during turns.

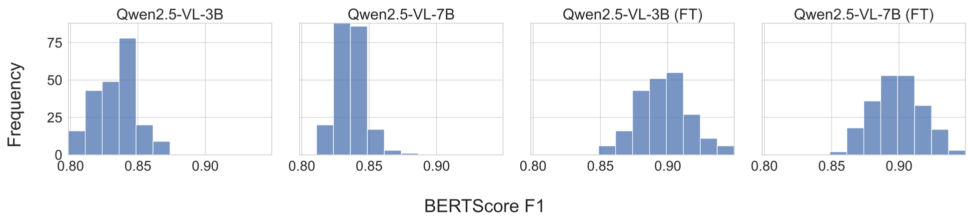

Figure 3: Score distribution of BERTScore F1 of event detection.

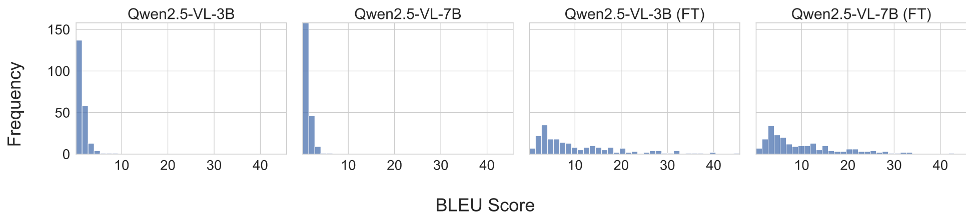

Figure 4: Score distribution of BLEU of event detection.

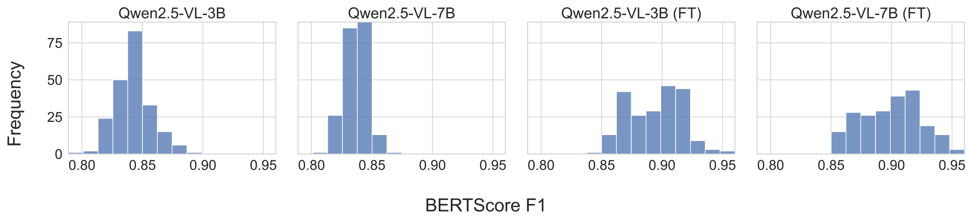

Figure 5: Score distribution of BERTScore F1 of safe driving instruction.

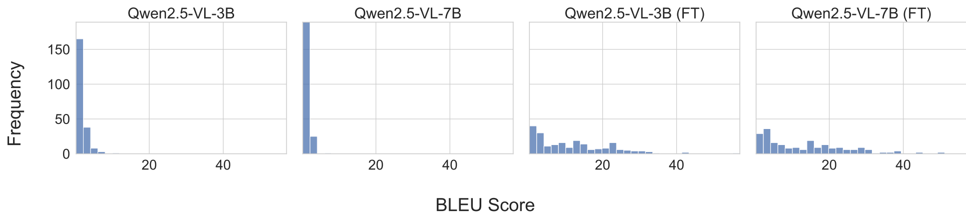

Figure 6: Score distribution of BLEU of safe driving instruction.

In the difficult subset, approximately 10% have errors related to the driver-facing view, where a driver holds and uses a phone while keeping their eyes on the road. This might suggest that LVLMs can provide safe driving instruction regarding drivers' behaviors, while they struggle to generate it regarding vehicle behaviors. This phenomenon is explained by the relative ease with which LVLMs can detect a driver holding an object, as opposed to estimating vehicle speed, which requires more temporal reasoning.

5.3 Unimodal Biases

As discussed in Section 5.1, in several cases, pretrained models provide general suggestions regardless of the videos, as shown in Table 8. We also observed that the fine-tuned models mentioned objects not presented in the video in Section 5.2. This behavior can be attributed to unimodal biases, specifically language biases (Goyal et al., 2017; Agrawal et al., 2018; Zhu et al., 2020; Abbasnejad et al., 2020; Chen et al., 2024), where models' outputs are biased toward textual information in the given inputs. However, while language biases have been attributed to the model's learning of the relationships between question-answer pairs in the training data (Agrawal et al., 2018), it is unclear whether the training data includes question-answer pairs regarding safe driving instruction. Given that Sakajo et al. (2025) observed the model generating identical answers for different images with the same question in a phonetics-related VQA task, our findings suggest that language priors emerge in certain domains.

5.4 Task Difficulty and Application Possibility

The results reveal that this task is challenging for LVLMs without fine-tuning, whereas fine-tuning improves performance. Our error analysis in Section 5.2 also indicates that several failure cases happen for good driving videos, and suggestions for safer driving vary slightly. Those discussions suggest that our fine-tuned models can be applied to safe driving instruction systems, although several challenges remain in certain situations, such as instructing against a harsh turn.

6 Conclusion

In this study, we constructed a dataset comprising synchronized driver-facing and road-facing video streams, along with step-by-step question–answer pairs. We fine-tuned LVLMs on our dataset for safe driving instructions and investigated their capabilities and performance in detecting risky events and providing safe driving instructions. Our experimental results reveal that fine-tuned LVLMs demonstrate the capability of suggesting safety-aware driving instructions, while detecting several events remains challenging, even for fine-tuned models. Our findings suggest that LVLMs can be safe driving instructors, although there is room for improvement.

Limitations

Dataset size. As described in Section 3, our dataset comprises 1,719 training samples, 215 validation samples, and 215 test samples, which can be considered relatively small in scale. However, the collection of synchronized driver-facing and road-facing views requires a complicated setup, characterizing this task as a low-resource scenario. In this study, we investigated model performance using the current dataset as an initial step, with evaluation on a larger dataset left for future work.

Dataset quality. The instructions in our dataset were generated using GPT-4o, which might raise concerns regarding their quality. However, as detailed in Section 3, the generated texts were manually reviewed to ensure the quality.

Model selection. In this study, we selected two base models: Qwen2.5-VL-3B and Qwen2.5-VL-7B. While this choice might constrain our investigation of the scaling law in this task and performance variation across models, it remains justifiable. The Qwen2.5-VL series achieves advanced performance on various benchmarks, including Video-MME (Fu et al., 2024), and our objective is to evaluate model effectiveness for driving instruction. Accordingly, focusing on the Qwen2.5-VL series and its relatively small variants is appropriate for our investigation.

Ethical Considerations

Our dataset contains videos that capture drivers. We collect these videos legitimately and use them within the prescribed scope.

Acknowledgement

We thank the anonymous reviewers for their valuable comments and suggestions. This project is supported by Teatis inc. through the provision of the dataset and computational resources.

References

Ehsan Abbasnejad, Damien Teney, Amin Parvaneh, Javen Shi, and Anton van den Hengel. 2020. Counterfactual vision and language learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10041–10051.

Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. 2018. Don't just assume; look and answer: Overcoming priors for visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob L Menick, Sebastian Borgeaud, and 8 others. 2022. Flamingo: A visual language model for few-shot learning. In *Advances in Neural Information Processing Systems*, volume 35, pages 23716–23736. Curran Associates, Inc.

Hidehisa Arai, Keita Miwa, Kento Sasaki, Kohei Watanabe, Yu Yamaguchi, Shunsuke Aoki, and Issei Yamamoto. 2025. Covla: Comprehensive vision-language-action dataset for autonomous driving. In *Proceedings of the Winter Conference on Applications of Computer Vision (WACV)*, pages 1933–1943.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Meiqi Chen, Yixin Cao, Yan Zhang, and Chaochao Lu. 2024. Quantifying and mitigating unimodal biases in multimodal large language models: A causal perspective. In *Findings of the Association for Computational*

Linguistics: EMNLP 2024, pages 16449–16469, Miami, Florida, USA. Association for Computational Linguistics.

Zhizhao Duan, Hao Cheng, Duo Xu, Xi Wu, Xiangxie Zhang, Xi Ye, and Zhen Xie. 2024. Cityllava: Efficient fine-tuning for vlms in city scenario. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 7180–7189.

Chaoyou Fu, Yuhan Dai, Yondong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, and 1 others. 2024. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. *arXiv preprint arXiv:2405.21075*.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Kazuki Hayashi, Yusuke Sakai, Hidetaka Kamigaito, Katsuhiko Hayashi, and Taro Watanabe. 2024. Towards artwork explanation in large-scale vision language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 705–729, Bangkok, Thailand. Association for Computational Linguistics.

Takeshi Kojima, Shixiang (Shane) Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems*, volume 35, pages 22199–22213.

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. 2023a. LLaVA-med: Training a large language-and-vision assistant for biomedicine in one day. In *Advances in Neural Information Processing Systems*, volume 36, pages 28541–28564. Curran Associates, Inc.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023b. Blip-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, volume 36, pages 34892–34916. Curran Associates, Inc.

Rongxin Liu, Carter Zenke, Charlie Liu, Andrew Holmes, Patrick Thornton, and David J. Malan. 2024. Teaching cs50 with ai: Leveraging generative artificial intelligence in computer science education. In *Proceedings of the 55th ACM Technical Symposium on Computer Science Education V. 1*, SIGCSE 2024, page 750–756, New York, NY, USA. Association for Computing Machinery.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *Preprint*, arXiv:1907.11692.

Yuhang Lu, Yichen Yao, Jiadong Tu, Jiangnan Shao, Yuexin Ma, and Xinge Zhu. 2025. Can lvlms obtain a driver's license? A benchmark towards reliable AGI for autonomous driving. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(6):5838–5846.

OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, and 401 others. 2024. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

Shintaro Ozaki, Kazuki Hayashi, Yusuke Sakai, Hidetaka Kamigaito, Katsuhiko Hayashi, and Taro Watanabe. 2025. Towards cross-lingual explanation of artwork in large-scale vision language models. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 3773–3809, Albuquerque, New Mexico. Association for Computational Linguistics.

Ankit Pal and Malaikannan Sankarasubbu. 2024. Gemini goes to Med school: Exploring the capabilities of multimodal large language models on medical challenge problems & hallucinations. In *Proceedings of the 6th Clinical Natural Language Processing Workshop*, pages 21–46, Mexico City, Mexico. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. 2020. Zero: memory optimizations toward training trillion parameter models. In *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, SC '20. IEEE Press.

Baptiste Rozière, Jonas Gehring, Fabian Gloeckle, Sten Sootla, Itai Gat, Xiaoqing Ellen Tan, Yossi Adi, Jingyu Liu, Romain Sauvestre, Tal Remez, Jérémy Rapin, Artyom Kozhevnikov, Ivan Evtimov, Joanna Bitton, Manish Bhatt, Cristian Canton Ferrer, Aaron

Grattafiori, Wenhan Xiong, Alexandre Défossez, and 7 others. 2024. Code llama: Open foundation models for code. *Preprint*, arXiv:2308.12950.

Haruki Sakajo, Yusuke Sakai, Hidetaka Kamigaito, and Taro Watanabe. 2025. Tonguescape: Exploring language models understanding of vowel articulation. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 12605–12619, Albuquerque, New Mexico. Association for Computational Linguistics.

Shuo Wang, David C. Anastasiu, Zheng Tang, Ming-Ching Chang, Yue Yao, Liang Zheng, Mohammed Shaiqur Rahman, Meenakshi S. Arya, Anuj Sharma, Pranamesh Chakraborty, Sanjita Prajapati, Quan Kong, Norimasa Kobori, Munkhjargal Gochoo, Munkh-Erdene Otgonbold, Ganzorig Batnasan, Fady Alnajjar, Ping-Yang Chen, Jun-Wei Hsieh, and 5 others. 2024. The 8th AI City Challenge. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*.

Zhenhua Xu, Yujia Zhang, Enze Xie, Zhen Zhao, Yong Guo, Kwan-Yee K. Wong, Zhenguo Li, and Hengshuang Zhao. 2024. Drivegpt4: Interpretable end-to-end autonomous driving via large language model. *IEEE Robotics and Automation Letters*, 9(10):8186–8193.

Khai Trinh Xuan, Khoi Nguyen Nguyen, Bach Hoang Ngo, Vu Dinh Xuan, Minh-Hung An, and Quang-Vinh Dinh. 2024. Divide and conquer boosting for enhanced traffic safety description and analysis with large vision language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 7046–7055.

Qianqi Yan, Xuehai He, Xiang Yue, and Xin Eric Wang. 2024. Worse than Random? An Embarrassingly Simple Probing Evaluation of Large Multimodal Models in Medical VQA. In *GenAI for Health: Potential, Trust and Policy Compliance*.

Senqiao Yang, Jiaming Liu, Ray Zhang, Mingjie Pan, Zoey Guo, Xiaoqi Li, Zehui Chen, Peng Gao, Yandong Guo, and Shanghang Zhang. 2023. Lidar-llm: Exploring the potential of large language models for 3d lidar understanding. *Preprint*, arXiv:2312.14074.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, and Zheyan Luo. 2024. LlamaFactory: Unified efficient fine-tuning of 100+ language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 400–410, Bangkok, Thailand. Association for Computational Linguistics.

Xingcheng Zhou, Mingyu Liu, Ekim Yurtsever, Bare Luka Zagar, Walter Zimmer, Hu Cao, and Alois C. Knoll. 2024. Vision language models in autonomous driving: A survey and outlook. *IEEE Transactions on Intelligent Vehicles*, pages 1–20.

Xi Zhu, Zhendong Mao, Chunxiao Liu, Peng Zhang, Bin Wang, and Yongdong Zhang. 2020. Overcoming language priors with self-supervised learning for visual question answering. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, IJCAI-PRICAI-2020, page 1083–1089. International Joint Conferences on Artificial Intelligence Organization.

Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. 2018. Texygen: A benchmarking platform for text generation models. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, SIGIR '18, page 1097–1100, New York, NY, USA. Association for Computing Machinery.

A Dataset Statistics (Detail)

Table 11 shows the number of events or options in each dataset split.

	Train	Val.	Test
Primary Events			
Crash	3	0	0
Forward Collision Warning	41	6	7
Tailgating	90	21	7
Harsh Brake	253	20	26
Harsh Turn	15	4	2
Rolling Stop	308	42	42
Mobile Usage	87	14	11
Inattentive Driving	142	17	20
Lane Departure	0	0	0
Sub Events			
Lane Cut Off			
Improper use of turn signal	18	3	0
Proper use of turn signal	21	7	1
Lane Change			
Improper use of turn signal	11	0	1
Proper use of turn signal	84	11	9
To avoid primary event	21	2	3
Root cause of primary event	24	3	2
Turn (Other Vehicles)			
Proper use of turn signal	10	1	3
Improper use of turn signal	3	0	0
Turn (Ego Vehicle)			
Proper use of turn signal	114	15	20
Improper use of turn signal	102	18	10
Signs of Aggressive Reaction			
Vehicle maneuver	1	0	0
Aggressive language	18	1	2
Honk	9	2	0
None	634	72	85
Unknown (Dashcam Issue)	209	30	25
Signs of Distraction			
Smoking	26	3	5
Mobile phone	18	3	2
Playing with hair	2	1	0
Drinking	14	0	2
Eating	18	2	3
Picking something from floor	2	0	1

Reaching behind backseat	0	1	0
Yawning	1	0	0
None	580	63	75
Unknown	218	32	25
Weather Condition			
Clear	1380	184	172
Rainy	92	6	10
Foggy	0	1	0
Snowy	60	7	9
Road Condition			
Dry	1362	181	170
Wet	98	6	11
Icy	72	11	11
Visibility Condition			
Clear	0	0	0
Poor	0	0	0
Road Information			
Highway	893	124	102
Highway merge	22	3	2
Local Road	8	0	0
Intersection	348	28	46
3-Leg intersection	166	31	26
School zone	0	1	0
Construction Zone	10	3	2
Residential area	66	8	14
Rural roads	12	2	1
Tunnel	1	0	0
Pedestrian crossroad	22	2	1
Parking	27	2	6
Speed Management			
Decrease	37	5	1
Maintain	409	59	48
Increase	43	9	1

Table 11: The number of each event or option in each dataset split. Val. denotes the validation set.