Road Rage Reasoning with Vision-language Models (VLMs): Task Definition and Evaluation Dataset

YIBING WENG, University of Electronic Science and Technology of China, China
YU GU, University of Electronic Science and Technology of China, China
FUJI REN, University of Electronic Science and Technology of China, China

Road rage, triggered by driving-related stimuli such as traffic congestion and aggressive driving, poses a significant threat to road safety. Previous research on road rage regulation has primarily focused on response suppression, lacking proactive prevention capabilities. With the advent of Vision-Language Models (VLMs), it has become possible to reason about trigger events visually and then engage in dialog-based comforting before drivers' anger escalates. To this end, we propose the road rage reasoning task, along with a finely annotated test dataset and evaluation metrics, to assess the capabilities of current mainstream VLMs in scene understanding, event recognition, and road rage reasoning. The results indicate that current VLMs exhibit significant shortcomings in scene understanding within the visual modality, as well as in comprehending the spatial relationships between objects in the textual modality. Improving VLMs' performance in these areas will greatly benefit downstream tasks like antecedent-focused road rage regulation.

CCS Concepts: • **Computing methodologies → Scene understanding**; *Activity recognition and understanding*; Video summarization.

Additional Key Words and Phrases: Road Rage Reasoning, Vision Language Models, Task Definition, Evaluation Dataset

1 Introduction

Road rage is defined as intense behaviorally maladaptive anger triggered by driving-related stimuli like traffic jams and aggressive driving [1]. It often leads to impulsive behaviors including swearing, deliberate ramming, and in extreme cases, even shooting [2], posing a significant threat to road safety. It has attracted attention from various fields. In transportation, researchers mainly focus on designing scales to analyze the relationship between age, personality, and impulsive behavior [3–6]. In psychology, the generation mechanisms [7] and regulation strategies [8, 9] of road rage have received much attention. In computer science, research focuses more on driver anger detection [10].

As for the regulation of road rage, the focus of the aforementioned research lies on the anger emotion of drivers, and thus the resulting regulatory measures are primarily based on side-channels such as music [11], scent [12], and lighting [13] to suppress potential impulsive behaviors. In other words, these methods aim to calm drivers who are already in a state of anger, but lack the proactive regulation capability (e.g., antecedent-focused regulation) to intervene before anger forms. This situation is mainly caused by technical difficulties in two aspects: proactive regulation requires prompt and accurate understandings of the potential trigger events as well as natural interactive means (such as conversational soothing) to address these causes. With the recent success of VLMs in autonomous driving tasks (especially in understanding external driving environments) [14–16], it naturally raises the following question: ***Can current VLMs handle these difficulties to enable antecedent-focused regulation (see Fig. 1)?***

As a first step in answering this question, we propose a road rage reasoning task for VLMs to evaluate how well they can understand the trigger events (i.e., antecedents) of potential road rage. The task takes visual inputs from a dashcam and assesses the capabilities of a VLM in scene understanding, event recognition, and road rage reasoning. The output is in text format, which can benefit downstream tasks like dialogue-based comforting, as shown in Fig. 2).

Authors' Contact Information: Yibing Weng, School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu, China, 202411081538@std.uestc.edu.cn; Yu Gu, School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu, China, yugu.bruce@ieee.org; Fuji Ren, School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu, China, renfuji@uestc.edu.cn.

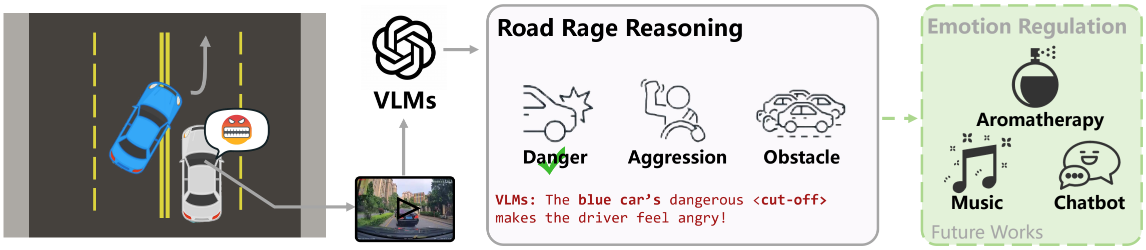


Fig. 1. We propose the road rage reasoning task to evaluate VLMs' capabilities in road rage visual understanding, behavior recognition, and scenario reasoning. Provides prior knowledge for downstream emotion regulation tasks, enabling antecedent-focused regulation.

To perform the task, we present a finely-annotated test dataset collected from real-world road rage dashcam footages along with various evaluation metrics. As shown in Fig. 3, the dataset consists of 81 videos, 2, 299 frames, and 22, 226 annotations, which cover both overall labels for one footage (environment descriptions, road rage events, and road rage scenarios) and detailed labels for each frame (lane count, ego car, and critical objects).

We tested two of the most widely-used VLMs, i.e., GPT-4o and Qwen-VL, on the task. The results revealed significant shortcomings in their visual scene understanding, which was crucial for event recognition. Furthermore, they struggled to comprehend spatial relationships between objects described in the text. For future work, we will try to fine-tune the VLMs with our dataset to enhance their capabilities in the above two aspects.

2 Related Works

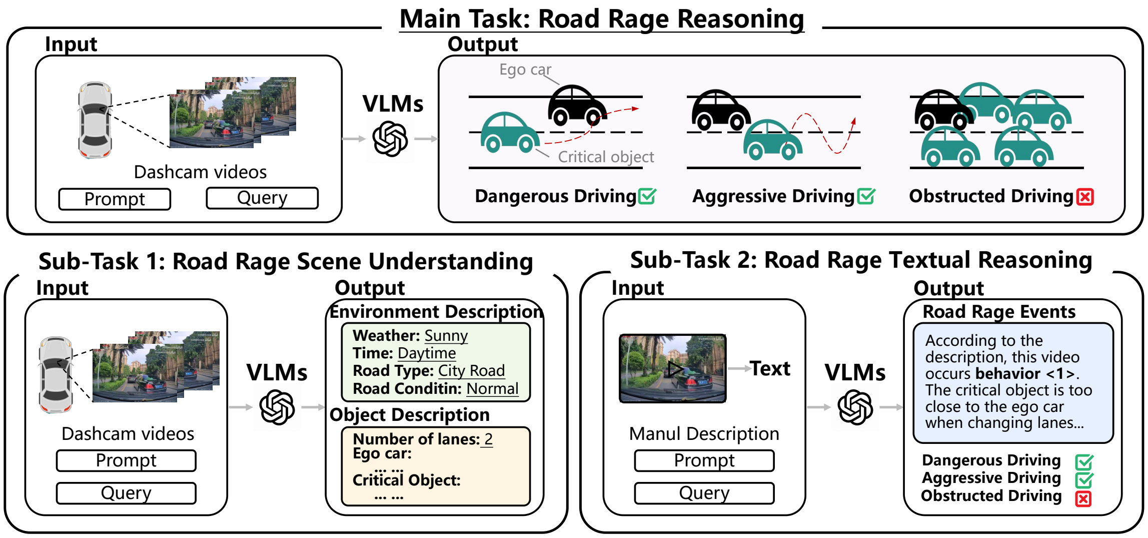
2.1 Road Rage

Currently, most studies on the causes of road rage use self-reports to understand which drivers are more prone to anger and which situations are more likely to trigger it [17]. These insights can be used for modeling and automatically predicting when a driver may become angry. Some methods have also created simulated driving scenarios to assess whether road rage occurs in different situations [18, 19]. These studies provide knowledge on the causes of road rage and suggest developing intervention strategies. Although some research proposes antecedent-focused regulation [20] to intervene before emotions occur, how to apply antecedent-focused regulation to road rage remains a challenge. This is because current techniques typically focus on detecting road rage after it has occurred, rather than identifying the causes that trigger it.

Common causes of road rage come from external driving environments [21]. Although some driving datasets have been introduced [22–24], which annotate objects in the environment and their potential threats to the ego car [25, 26], these datasets mainly focus on training autonomous driving systems rather than perceiving external factors which cause road rage. There is still a lack of datasets specifically focused on road rage in driving environments.

2.2 VLMs in Driving Tasks

VLMs have made significant progress in the field of autonomous driving [27]. By combining computer vision and natural language processing, VLMs can understand and process complex information encountered during driving, such as road signs [28], vehicle behavior, and pedestrian detection [29]. Numerous studies have shown that VLMs exhibit strong multimodal understanding abilities in driving environments, allowing them to comprehend both visual and textual inputs and make accurate inferences.



According to the description, this video occurs **behavior <1>**. The critical object is too close to the ego car when changing lanes...

Fig. 2. We design three tasks to evaluate VLMs' road rage reasoning abilities. The **Main Task** uses dashcam footage to identify road rage scenarios, testing overall reasoning. Due to poor performance, we introduce two subtasks. **Sub-task 1** uses dashcam footage to assess scene understanding but lacks complete responses for quantitative analysis. Thus, **Sub-task 2** uses manual descriptions, decoupling visual understanding from reasoning, and assesses textual reasoning and scene understanding capabilities.

In recent studies ([14, 16, 30]), VLMs have demonstrated impressive performance in driving and driving-related dialogue tasks, including visual understanding of the external environment, visual question answering based on that understanding, and the ability to infer the potential impact of external objects on the current vehicle. This shows the potential of VLMs to reason about the external driving environment that may trigger road rage.

3 Task Definition and Test Dataset Collection

In this section, we define a main task and two sub-tasks to test VLMs' capabilities in road rage reasoning, (as shown in Fig. 2) and explain the rationale behind the design of these tasks. We then provided a detailed description of the dataset collection and annotation process, including the criteria for video collection and the content of the annotations.

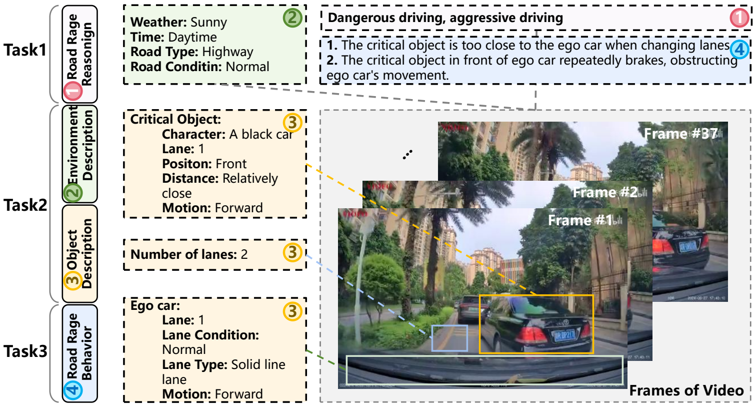
3.1 Task Definition

3.1.1 Main task: road rage reasoning. The main task requires VLMs to understand the dashcam footage as a whole and determine whether dangerous, aggressive, or obstructive driving are present. This task implicitly demands full scene understanding and the ability to capture key features for situation assessment. Additionally, VLMs must recognize events in context and link them to corresponding road rage scenarios. Due to poor performance in the main task, we introduce two sub-tasks to explore the factors limiting VLMs' performance in road rage reasoning tasks.

3.1.2 Sub-task1: road rage scene understanding. Sub-task 1 also uses dashcam videos, where VLMs are required to provide detailed descriptions of each frame based on a template we provide. The responses are in a fill-in-the-blank format, including environment and object descriptions (see Fig. 3 (b)). The template defines options for each blank to aid subsequent calculations. Unfortunately, despite testing various prompts, VLMs still fail to describe each frame, preventing us from drawing quantifiable conclusions. Therefore, we set up sub-task 2.

Video selection	
Main source	www.bilibili.com
Annotation language	en / ch
Number of videos	81
Sampling rate	2 fps
Total frames	2299
Resolution	512 × 288 pixels
Number of frames	15 - 37
Details	
Weather	3
Road types	3
Road conditions	3
Ego car's actions	4
Critical object's actions	10
Annotations	
Road rage scenarios	81 × 3
Road rage events	81 × 9
Environment descriptions	81 × 4
Number of lanes	2299 × 1
Ego car annotations	2299 × 4
Critical object annotations	1887 × 5
Total	22226

(a) Dataset statistic



(b) Dataset labeling sample

Fig. 3. The statistics (a) and an annotation example (b) of our dataset. The dataset includes 81 videos, 2,299 frames, and 22,226 annotations. The annotations cover both overall labels (environment descriptions, road rage events and road rage scenarios) and detailed labels (lane count, ego car, and critical objects).

3.1.3 Sub-task2: road rage textual reasoning. Sub-task 2 converts the video labels into manual descriptions and uses them as input for VLMs. VLMs are then asked to recognize road rage events. Based on the output events of the VLMs, we categorize the road rage scenarios. VLMs can extract all relevant information related to road rage scenarios from manual descriptions. Therefore, by using this method, we can assume that VLMs have perfect scene understanding, decoupling scene understanding from reasoning. The conclusions from the experiment directly reflect VLMs' performance in textual reasoning and indirectly highlight their shortcomings in scene understanding.

3.2 Dataset Collection

Fig. 3 is an overview of the dataset. It shows the number of videos we collected, the number of frames divided, the total number of annotations, and provides an example of video annotation.

3.2.1 Selection criteria. To create the road rage reasoning dataset, we recruit three volunteers, each with at least two years of driving experience. They manually select videos from popular platforms such as YouTube, Bilibili, and Youku. The main criterion for video selection is: "Does the video make you feel angry, anxious, or tense, and does it show at least one of the following: dangerous driving, aggressive driving, or obstructive driving?". Additionally, to standardize the content, the following restrictions are applied:

- The video must be recorded from the driver's first-person perspective.
- The video should not contain commentary subtitles.
- The events in the video must be easily observable and describable.
- The video must cover the entire duration of the event.
- The videos must be from the same region, as driving rules may vary by location.

Through this process, over 100 driving videos are collected, each initially annotated with a road rage scenario classification label.

Table 1. The number of specific events in the collected videos of the three road rage scenarios.

Type	Event Description	Number
Dangerous Driving	1. The critical object is too close to the ego car when changing lanes.	44
	2. The critical object did not use the turn signal when changing lanes.	27
	3. The critical object involves illegally changing lanes across a solid line while driving.	14
	4. The critical object is too close to the ego car when overtaking the ego car.	16
	5. The critical object is to brake, decelerate, or stop suddenly at a location very close to the ego car.	39
	6. The critical object is a non-vehicle entity (such as a pedestrian, electric scooter, bicycle, etc.) crossing the road.	11
Aggressive Driving	7. The critical object in front of ego car repeatedly brakes, obstructing ego car's movement.	19
	8. The critical object repeatedly cuts into ego car's lane, displaying aggressive driving behavior.	13
Obstructive Driving	9. Roadways are generally busy or congested.	32

3.2.2 Road rage events labeling. In our approach, we focus on having VLMs recognize the specific events that provoke anger, rather than just classifying scenarios into general categories. To achieve this, we annotate the road rage events in the collected videos. To ensure the events in the filtered videos are sufficiently anger-inducing, we refer to the MAD [31]. The MAD scale assesses the level of anger triggered by different driving events. Based on the anger scores from MAD, we recalculate rankings using a weighted average method.

Additionally, since some events in the MAD scale conflict with the rules for the videos we collected, we replace these events accordingly. After this, we apply secondary filtering to the videos based on two criteria: first, the anger score induced by the event should be as high as possible, and second, each event should have at least 10 samples.

3.2.3 Scene labeling. The next step involves detailed annotation of the road conditions, the behaviors of the ego car, and the behaviors of the critical objects involved in the road rage scenario. For the entire video, we first provide environment descriptions, including weather, time, road type, and road conditions. Then, we sample the video at a rate of two frames per second. For each sampled frame, we perform object descriptions, including the number of lanes on the current road, as well as the positions and behaviors of the ego car and critical objects. For the ego car, we describe its lane, the types of lane markings on both sides of the lane, and its actions (the actions are based on the DriveVLM [16]). For the critical object, we first describe its visual features, followed by its lane, its relative position to the ego car, the distance between them, and its actions. To facilitate subsequent calculations, we set predefined options for each variable. Fig. 3 (b) shows an example of our annotation.

3.2.4 Dataset statistics. After filtering, we ultimately retain 81 videos. Based on the time range in which road rage events occur, and after sampling two frames per second, we obtain a total of 2,299 frames. We perform both overall annotations for the videos and detailed per-frame annotations, resulting in a total of 22,226 annotations. Fig. 3 (a) provides a reference. In Table 1, we list all the road rage events that occur in the videos, along with their frequency (note that multiple events can appear in a single video). These events correspond to specific road rage scenarios. Among them, dangerous driving includes 6 types of behaviors: 5 related to dangerous or illegal actions of vehicles, and 1 related to non-vehicle objects crossing the road. Aggressive driving consists of 2 types of behaviors: repeatedly braking to block the ego car and repeatedly cutting into the ego car's lane. Obstructive driving includes 1 situation: road congestion or heavy traffic.

Table 2. The performance of GPT-4o and Qwen-VL on the main task. Bold text and underlining represent the maximum and minimum values of precision and recall, respectively. VLMs perform best in obstructive driving, which is more visually-oriented, and worst in aggressive driving, which requires more reasoning ability.

Models	Dangerous driving		Aggressive driving		Obstructed Driving		Accuracy
	Precision	Recall	Precision	Recall	Precision	Recall	
GPT-4o	**100.0**	31.17	<u>12.50</u>	<u>12.50</u>	61.11	**68.75**	14.81
Qwen-VL	**100.0**	28.57	<u>50.00</u>	<u>4.17</u>	62.79	**84.38**	18.52

4 Evaluation and Results

We select gpt-4o-2024-08-06 and qwen-vl-max-2024-11-19 as baseline models. All experiments are conducted in a zero-shot setting. The evaluation metrics include precision, recall, and accuracy, which help us assess VLMs' misjudgment rate, their ability to correctly identify specific scenarios or events, and their overall reasoning capability.

4.1 Evaluation of Main Task

Firstly, we present the results of the main task, as shown in Table 2. These results reflect VLMs' ability in overall reasoning. Based on the data from the table, we summarize our findings as follows:

VLMs show significant room for improvement in reasoning road rage scenarios: We observe that both GPT-4o and Qwen-VL show low accuracy in reasoning road rage based on the whole video, indicating that VLMs can't effectively understand video content.

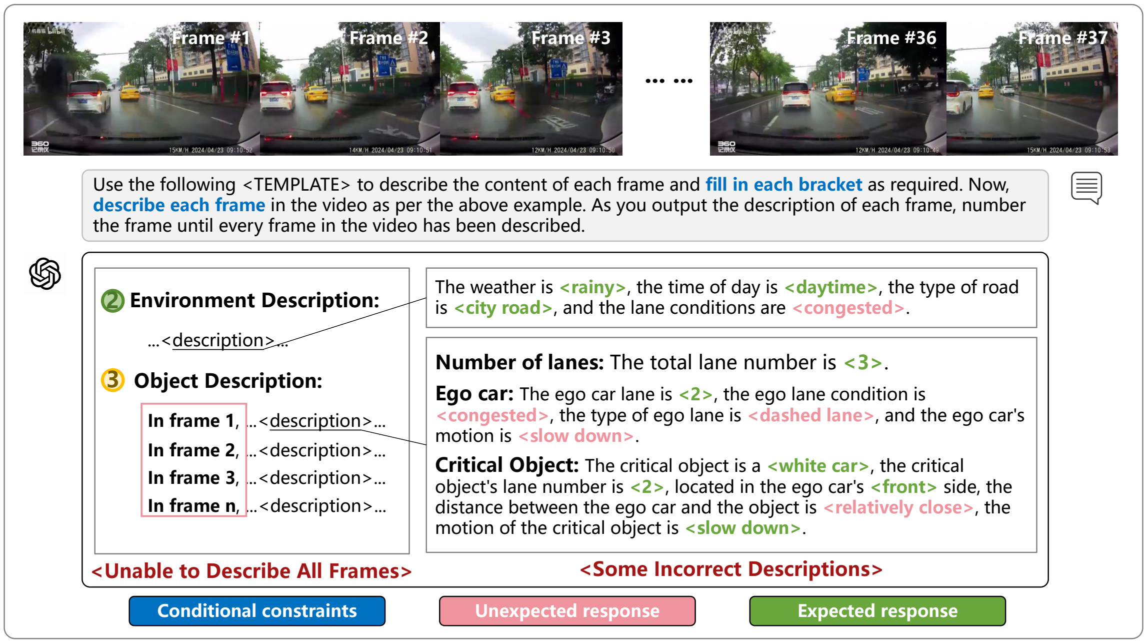
VLMs exhibit significant performance differences across different road rage scenarios: The recall rates show that Qwen-VL performs well for visually-oriented obstructive driving, while GPT-4o performs slightly worse. For dangerous driving, VLMs can partially classify the scenarios. However, for reasoning-oriented aggressive driving, VLMs show the worst recall rates. This makes sense because obstructive driving is easier to detect due to the presence of many surrounding vehicles. Most dangerous driving incidents in our dataset involve critical object being very close to ego car, which makes it easier for VLMs to identify if VLMs can capture this feature. Aggressive driving, however, requires recognizing the events first and then inferring its aggressive nature, which is a significant challenge for current VLMs.

VLMs tend to make misjudgments: This conclusion is derived from comparing precision values. For dangerous driving, while recall rates show that not all dangerous behaviors are recognized, the precision is high. This indicates that VLMs make some correct judgments. For obstructive driving, while VLMs demonstrate relatively good recall, precision shows that there are still many misjudgments. For aggressive driving, precision is still poor.

4.2 Evaluation of Sub-task 1

Sub-task 1 is a scene understanding task. In the prompt, we provide constraints and ask VLMs to fill in the blanks based on the template. VLMs must also describe each frame of the input. Fig. 4 shows an example answer for sub-task 1.

The answers consist of two parts: environment description and object description (see Section 3.2). From the object description, we see that, despite the constraints, VLMs still cannot describe every frame. Additionally, there are some incorrect descriptions in both environment and object details. These results show that VLMs have limitations in scene understanding. Since VLMs cannot fully answer the question, we also cannot perform a quantitative analysis or draw firm conclusions. Therefore, we conduct additional experiments (Sub-task 2) to further explore VLMs' understanding and reasoning abilities.


Frame #1
Frame #2
Frame #3
Frame #36
Frame #37
② Environment Description:
...<description>...
③ Object Description:
In frame 1, ...<description>...
In frame 2, ...<description>...
In frame 3, ...<description>...
In frame n, ...<description>...
<Unable to Describe All Frames>
<Some Incorrect Descriptions>
Conditional constraints
Unexpected response
Expected response

Use the following <TEMPLATE> to describe the content of each frame and fill in each bracket as required. Now, describe each frame in the video as per the above example. As you output the description of each frame, number the frame until every frame in the video has been described.

The weather is <rainy>, the time of day is <daytime>, the type of road is <city road>, and the lane conditions are <congested>.

Number of lanes: The total lane number is <3>.

Ego car: The ego car lane is <2>, the ego lane condition is <congested>, the type of ego lane is <dashed lane>, and the ego car's motion is <slow down>.

Critical Object: The critical object is a <white car>, the critical object's lane number is <2>, located in the ego car's <front> side, the distance between the ego car and the object is <relatively close>, the motion of the critical object is <slow down>.

Fig. 4. An experimental result from sub-task 1. Under the given constraints, VLMs still fail to describe all frames. In the details, VLMs show some incorrect descriptions. This result prevents us from performing a quantitative analysis of VLMs' visual understanding ability. Therefore, we introduce sub-task 2.

Table 3. The experimental results of sub-task 2, where the underlined values represent the minimum. Using manual descriptions as input, VLMs can accurately reason most road rage events. However, there is still room for improvement in events requiring spatial relationship comprehension (Events 3, 7, and 8).

Models	Road rage scenarios	Dangerous driving						Aggressive driving		Obstructive driving
	Road rage events	1	2	3	4	5	6	7	8	9
GPT-4o	Precision	87.76	81.82	77.78	88.89	90.70	100.0	45.24	58.82	100.0
	Recall	97.73	100.0	50.00	100.0	100.0	100.0	100.0	76.92	100.0
	Accuracy	39.51								
Qwen-VL	Precision	83.02	67.50	68.75	84.21	86.67	100.0	46.15	31.58	100.0
	Recall	100.0	100.0	78.57	100.0	100.0	100.0	94.74	92.31	96.88
	Accuracy	29.63								

Table 4. The road rage scenarios derived from the events output by sub-task 2, with the results from VLMs based on video input shown after the slash. Except for a slight decrease in precision for some scenarios, other metrics have shown significant improvement.

Models	Dangerous driving		Aggressive driving		Obstructed Driving		Accuracy
	Precision	Recall	Precision	Recall	Precision	Recall	
GPT-4o	96.25 / 100.0	100.0 / 31.17 ↑	46.00 / 12.50 ↑	95.83 / 12.50 ↑	100.0 / 61.11 ↑	100.0 / 68.75 ↑	62.96 / 14.81 ↑
Qwen-VL	96.25 / 100.0	100.0 / 28.57 ↑	39.66 / 50.00	95.83 / 4.17 ↑	100.0 / 62.79 ↑	100.0 / 84.38 ↑	53.09 / 18.52 ↑

4.3 Evaluation of Sub-task 2

Sub-task 2 converts video inputs into manual descriptions, assuming VLMs have perfect scene understanding, thereby decoupling understanding from reasoning. Table 3 presents the results of recognition of road rage events, and Table 4 shows the classification outcomes of road rage scenarios from these events. We summarize our findings as follows:

VLMs' scene understanding limits road rage reasoning: Table 3 demonstrates that VLMs recognize most events accurately and achieve high precision and recall when using manual descriptions as input. Table 4 compares road rage scenario classification results using video and textual modalities. Except for a slight drop in precision for some scenarios, other metrics show significant improvement. This highlights that limitations in scene understanding restrict VLMs' ability to reason road rage events and scenarios.

VLMs' spatial relationship comprehension limits textual reasoning: Table 3 reveals that precision for aggressive driving events (Events 7 and 8) remains low. Additionally, the recall rate for Event 3 (critical objects crossing solid lines to overtake) has room for improvement. These results indicate that, even with complete scene information, VLMs struggle with tasks requiring spatial relationship comprehension, particularly for Events 3, 7, and 8.

VLMs still perform misjudgment: Despite significant improvements in road rage events recognition and scenarios classification when converting the visual modality to the manual description, VLMs still perform event misjudgments. Specifically, the precision for certain events, particularly aggressive driving events, remains low. This indicates that many events are mistakenly classified as aggressive.

5 Discussion and Future Works

5.1 Discussion

Based on the above experiments, we summarize two key observations: 1. VLMs' limitations in scene understanding for road rage reasoning. While VLMs have made progress in identifying visual features, their ability to understand and reason about complex events remains a challenge. 2. VLMs' shortcomings in spatial relationship comprehension with textual modality. Converting video inputs into textual modality assumes perfect scene understanding, yet VLMs struggle with spatial relationship comprehension tasks. Then, we propose two open questions:

1. How can VLMs' understanding of complex scenes be enhanced? Despite advances in visual feature recognition, enabling VLMs to understand and reason more intricate events remains a challenge. Could stronger reasoning modules or optimized training and fine-tuning strategies address this issue in future research?

2. How can VLMs' spatial relationship comprehension in the text modality be improved? Even with perfect scene understanding through textual inputs, VLMs exhibit significant weaknesses in spatial relationship comprehension, particularly in tasks that involve cross-frame comprehension of vehicle positions and movements. How can we enhance VLMs' capabilities in spatial relationship comprehending within the text modality, especially for tasks requiring multi-frame understanding?

5.2 Future Works

Based on the two questions raised above, we discuss feasible solutions as future work:

1. Fine-tuning VLMs using our collected dataset or existing datasets. Existing studies and our experimental results indicate that current VLMs perform poorly in specific tasks like driving. The likely reasons are a lack of driving-related content in the training data and insufficient understanding of movements. Therefore, future work could involve fine-tuning VLMs using driving datasets (including image and action annotations) through instruction-based fine-tuning or full parameter fine-tuning. Given the lack of action annotations in existing driving datasets, a potential solution might be to fine-tune action understanding datasets first, then transfer to driving tasks. Alternatively, deep learning-based methods, which excel in object recognition and action understanding, could be explored.

2. Fine-tuning VLMs for spatial relationship understanding in textual modality. Most VLMs use the CLIP architecture, which connects visuals to text. Improving VLMs' understanding of spatial relationships in text can also boost their visual reasoning. To do this, future work can add special training for spatial relationships in textual reasoning. Using data that focuses on spatial relationships (e.g., object positions, speed changes) can guide VLMs to learn how to infer spatial changes in different situations. Reinforcement learning methods could be used to improve the model's ability to understand spatial relationships across frames.

6 Conclusion

We propose the road rage reasoning tasks and evaluate VLMs' scene understanding and textual reasoning abilities. We set up three tasks around road rage reasoning. The main task tests VLMs' overall reasoning ability, but the performance is poor. Sub-task 1 tests VLMs' scene understanding, but they fail to provide complete answers. Sub-task 2 converts the video into manual descriptions, decoupling visual understanding from reasoning. The results show that VLMs have limitations in scene understanding and spatial relationship comprehension in road rage reasoning. Therefore, we raise open questions, hoping to solve these two issues. Solving these problems can provide prior knowledge for downstream tasks like emotion regulation, contributing to a safer and more comfortable driving environment.

References

[1] Tara E Galovski, Loretta S Malta, and Edward B Blanchard. 2006. *Road rage: assessment and treatment of the angry, aggressive driver.* American Psychological Association. https://doi.org/10.1037/11297-000

[2] Elisabeth Wells-Parker, Jennifer Ceminsky, Victoria Hallberg, Ronald W Snow, Gregory Dunaway, Shawn Guiling, Marsha Williams, and Bradley Anderson. 2002. An exploratory study of the relationship between road rage and crash experience in a representative sample of US drivers. *Accident Analysis & Prevention* 34, 3 (2002), 271–278. https://doi.org/10.1016/S0001-4575(01)00021-5

[3] José María Faílde-Garrido, Yolanda Rodríguez-Castro, Antonio González-Fernández, and Manuel Antonio García-Rodríguez. 2023. Traffic Crimes and risky driving: The role of personality and driving anger. *Current Psychology* 42, 14 (2023), 12281–12295. https://doi.org/10.1007/s12144-021-02634-2

[4] Ning An, Long Sun, and Zhitai Wei. 2023. Adaptation and validity of the reckless driving habits scale in young Chinese drivers. *Transportation Research Part F: Traffic Psychology and Behaviour* 93 (2023), 174–181. https://doi.org/10.1016/j.trf.2023.01.011

[5] Dalal Youssef, Pascale Salameh, Linda Abou-Abbas, and Louis-Rachid Salmi. 2023. Driving anger dimensions and their relationship with aberrant driver behavior in Lebanon: Results from a national self-reported survey. *PLOS ONE* 18, 3 (2023), 1–17. https://doi.org/10.1371/journal.pone.0283293

[6] İbrahim Öztürk, Pınar Bıçaksız, Yeşim Üzümcüoğlu, and Türker Özkan. 2024. MAD-DAX: Anger and driver behaviours on the road. *Transportation Research Part F: Traffic Psychology and Behaviour* 102 (2024), 107–124. https://doi.org/10.1016/j.trf.2024.02.016

[7] Johan Bjureberg and James J. Gross. 2021. Regulating road rage. *Social and Personality Psychology Compass* 15, 3 (2021), e12586. https://doi.org/10.1111/spc3.12586

[8] Weina Qu, Muyi Liu, and Yan Ge. 2024. Mindfulness decreases driving anger expression: The mediating effect of driving anger and anger rumination. *Accident Analysis & Prevention* 203 (2024), 107642. https://doi.org/10.1016/j.aap.2024.107642

[9] Steven Love, Lee Kannis-Dymand, Grégoire S. Larue, and Bevan Rowland. 2024. Development and validation of the beliefs about driver anger questionnaire: A scale to predict anger propensity on the road. *Transportation Research Part F: Traffic Psychology and Behaviour* 102 (2024), 1–15. https://doi.org/10.1016/j.trf.2024.02.002

[10] Sebastian Zepf, Javier Hernandez, Alexander Schmitt, Wolfgang Minker, and Rosalind W. Picard. 2020. Driver Emotion Recognition for Intelligent Vehicles: A Survey. *ACM Comput. Surv.* 53, 3 (2020), 1–30. https://doi.org/10.1145/3388790

[11] Seyedeh Maryam Fakhrhosseini, Steven Landry, Yin Yin Tan, Saru Bhattarai, and Myounghoon Jeon. 2014. If You're Angry, Turn the Music on: Music Can Mitigate Anger Effects on Driving Performance. In *Proceedings of the 6th International Conference on Automotive User Interfaces and Interactive Vehicular Applications* (Seattle, WA, USA) *(AutomotiveUI '14)*. Association for Computing Machinery, New York, NY, USA, 1–7. https://doi.org/10.1145/2667317.2667410

[12] Dmitrijs Dmitrenko, Emanuela Maggioni, Giada Brianza, Brittany E. Holthausen, Bruce N. Walker, and Marianna Obrist. 2020. CARoma Therapy: Pleasant Scents Promote Safer Driving, Better Mood, and Improved Well-Being in Angry Drivers. In *Proceedings of the 2020 CHI Conference*

on Human Factors in Computing Systems (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3313831.3376176

[13] Mariam Hassib, Michael Braun, Bastian Pfleging, and Florian Alt. 2019. Detecting and Influencing Driver Emotions Using Psycho-Physiological Sensors and Ambient Light. In *Human-Computer Interaction – INTERACT 2019*, David Lamas, Fernando Loizides, Lennart Nacke, Helen Petrie, Marco Winckler, and Panayiotis Zaphiris (Eds.). Springer International Publishing, Cham, 721–742. https://doi.org/10.1007/9

[14] Chonghao Sima, Katrin Renz, Kashyap Chitta, Li Chen, Hanxue Zhang, Chengen Xie, Jens Beißwenger, Ping Luo, Andreas Geiger, and Hongyang Li. 2025. DriveLM: Driving with Graph Visual Question Answering. In *Computer Vision – ECCV 2024*. Springer, Springer Nature Switzerland, Cham, 256–274. https://doi.org/10.1007/978-3-031-72943-0_15

[15] Yingzi Ma, Yulong Cao, Jiachen Sun, Marco Pavone, and Chaowei Xiao. 2025. Dolphins: Multimodal Language Model for Driving. In *Computer Vision – ECCV 2024*. Springer, Springer Nature Switzerland, Cham, 403–420. https://doi.org/10.1007/978-3-031-72995-9_23

[16] Xiaoyu Tian, Junru Gu, Bailin Li, Yicheng Liu, Yang Wang, Zhiyong Zhao, Kun Zhan, Peng Jia, Xianpeng Lang, and Hang Zhao. 2024. DriveVLM: The Convergence of Autonomous Driving and Large Vision-Language Models. arXiv:2402.12289 [cs.CV] https://arxiv.org/abs/2402.12289

[17] Līva Ābele, Sonja Haustein, Mette Møller, and Ingo Zettler. 2020. Links between observed and self-reported driving anger, observed and self-reported aggressive driving, and personality traits. *Accident Analysis & Prevention* 140 (2020), 105516. https://doi.org/10.1016/j.aap.2020.105516

[18] Yi Wang, Jingyue Zheng, Wei Zhang, Xing Chen, and Min Wu. 2024. Inducing driving anger with multi-stage road events in simulator environment. *Transportation Research Part F: Traffic Psychology and Behaviour* 101 (2024), 403–422. https://doi.org/10.1016/j.trf.2024.02.001

[19] Myeongkyu Lee, Sangho Lee, Sungwook Hwang, Sejoon Lim, and Ji Hyun Yang. 2023. Effect of emotion on galvanic skin response and vehicle control data during simulated driving. *Transportation Research Part F: Traffic Psychology and Behaviour* 93 (2023), 90–105. https://doi.org/10.1016/j.trf.2022.12.010

[20] James J Gross. 1998. Antecedent-and response-focused emotion regulation: divergent consequences for experience, expression, and physiology. *Journal of personality and social psychology* 74, 1 (1998), 224–237. https://doi.org/10.1037/0022-3514.74.1.224

[21] Jerry L. Deffenbacher. 2016. A review of interventions for the reduction of driving anger. *Transportation Research Part F: Traffic Psychology and Behaviour* 42 (2016), 411–421. https://doi.org/10.1016/j.trf.2015.10.024

[22] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. 2020. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 11621–11631.

[23] Xinyu Huang, Xinjing Cheng, Qichuan Geng, Binbin Cao, Dingfu Zhou, Peng Wang, Yuanqing Lin, and Ruigang Yang. 2018. The apolloscape dataset for autonomous driving. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. 954–960.

[24] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, Vijay Vasudevan, Wei Han, Jiquan Ngiam, Hang Zhao, Aleksei Timofeev, Scott Ettinger, Maxim Krivokon, Amy Gao, Aditya Joshi, Yu Zhang, Jonathon Shlens, Zhifeng Chen, and Dragomir Anguelov. 2020. Scalability in Perception for Autonomous Driving: Waymo Open Dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2446–2454.

[25] Gurkirt Singh, Stephen Akrigg, Manuele Di Maio, Valentina Fontana, Reza Javanmard Alitappeh, Salman Khan, Suman Saha, Kossar Jeddisaravi, Farzad Yousefi, Jacob Culley, Tom Nicholson, Jordan Omokeowa, Stanislao Grazioso, Andrew Bradley, Giuseppe Di Gironimo, and Fabio Cuzzolin. 2023. ROAD: The Road Event Awareness Dataset for Autonomous Driving. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45, 1 (2023), 1036–1054. https://doi.org/10.1109/TPAMI.2022.3150906

[26] Scott Ettinger, Shuyang Cheng, Benjamin Caine, Chenxi Liu, Hang Zhao, Sabeek Pradhan, Yuning Chai, Ben Sapp, Charles R. Qi, Yin Zhou, Zoey Yang, Aurélien Chouard, Pei Sun, Jiquan Ngiam, Vijay Vasudevan, Alexander McCauley, Jonathon Shlens, and Dragomir Anguelov. 2021. Large Scale Interactive Motion Forecasting for Autonomous Driving: The Waymo Open Motion Dataset. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 9710–9719.

[27] 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* (2024), 1–20. https://doi.org/10.1109/TIV.2024.3402136

[28] Xu Cao, Tong Zhou, Yunsheng Ma, Wenqian Ye, Can Cui, Kun Tang, Zhipeng Cao, Kaizhao Liang, Ziran Wang, James M. Rehg, and Chao Zheng. 2024. MAPLM: A Real-World Large-Scale Vision-Language Benchmark for Map and Traffic Scene Understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 21819–21830.

[29] 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*. 7180–7189.

[30] Long Chen, Oleg Sinavski, Jan Hünermann, Alice Karnsund, Andrew James Willmott, Danny Birch, Daniel Maund, and Jamie Shotton. 2024. Driving with LLMs: Fusing Object-Level Vector Modality for Explainable Autonomous Driving. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*. 14093–14100. https://doi.org/10.1109/ICRA57147.2024.10611018

[31] A.N. Stephens, A. Lennon, C. Bihler, and S. Trawley. 2019. The measure for angry drivers (MAD). *Transportation Research Part F: Traffic Psychology and Behaviour* 64 (2019), 472–484. https://doi.org/10.1016/j.trf.2019.06.002

A Dataset Details

To facilitate the analysis of VLMs' responses, we define a set of possible options for each labeled variable. VLMs can only choose from these options, as shown in Table 5. It's important to note that the actions of critical objects are often complex and varied. To address this, we categorized the actions into general actions, vehicle actions, vehicle signals, and other actions. For vehicle objects, we use general actions, vehicle actions, and vehicle signals. For non-vehicle objects, we use general actions and other actions to describe their behavior.

Table 5. The details of environment description and object description, including the selectable options for each variable.

Descriptions	Variable	Options	
Environment Description	weather	sunny/rainy/unknow	
	time	daytime/night /unknow	
	road type	highway/city road/township road	
	road condition	normal /busy/congested	
Object Description	number of lanes	integer	
	ego lane number	integer	
	ego lane condiiton	normal/congested	
	ego lane type	dashed lane/solid lane/dashed-solid lane/solid-dashed lane	
	ego actions	forward/stop/go ahead to the left/go ahead to the right	
	critical object description	like white car, red van, electric car, pedestrian and so on	
	critical object lane number	integer	
	critical object relative position	left/right/front/left front/right front	
	critical object distance	normal/relatively close/dangerous	
	critical object actions	general actions	forward/slow down/stop/go ahead to the left/go ahead to the right
		vehicle actions	overtake/pull into ego lane to the left/ pull into ego lane to the right
		vehicle signals	no turn signal/brake
		other actions	crossing the street from left to right/ crossing the street from right to left

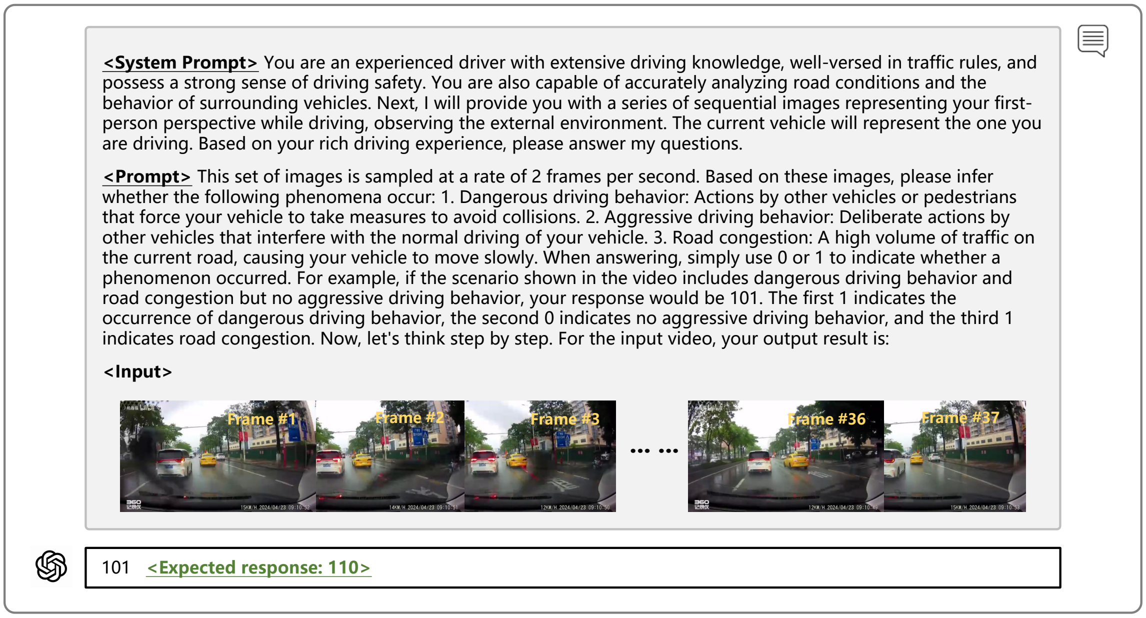
B Experiment Results

Fig. 5, 6, 7 provide examples from the experiments of the main task, Sub-task 1, and Sub-task 2. These include how we set the prompts, the input formats, the output constraints, and the complete responses from the VLMs.

We also provide two experimental cases to explain our observations, as shown in Fig. 8. For scene understanding, we ask VLMs to select key frames based on prompts (e.g., when the critical object is too close to the ego vehicle). VLMs fail to judge distance visually and miss some key frames. For textual reasoning, we ask if event 8 (critical object repeatedly cuts into ego car's lane) appears based on manual descriptions. VLMs fail to deduce the correct conclusion from spatial changes, showing irrelevant reasoning. This suggests that VLMs need to improve the understanding of spatial relationships in text.

C Exploration of Future Works

Our future work will first focus on fine-tuning VLMs to enable them to perceive external driving environments and reason about road rage. Then, we will explore emotional regulation methods, such as comforting conversations, aromatherapy, and music-based interventions. As shown in Fig. 9. The ideal scenario is for VLMs to predict potential causes of road rage before it occurs, and to implement targeted strategies using emotional regulation techniques to alleviate the driver's road rage. This work could contribute to making future driving environments safer and more comfortable.



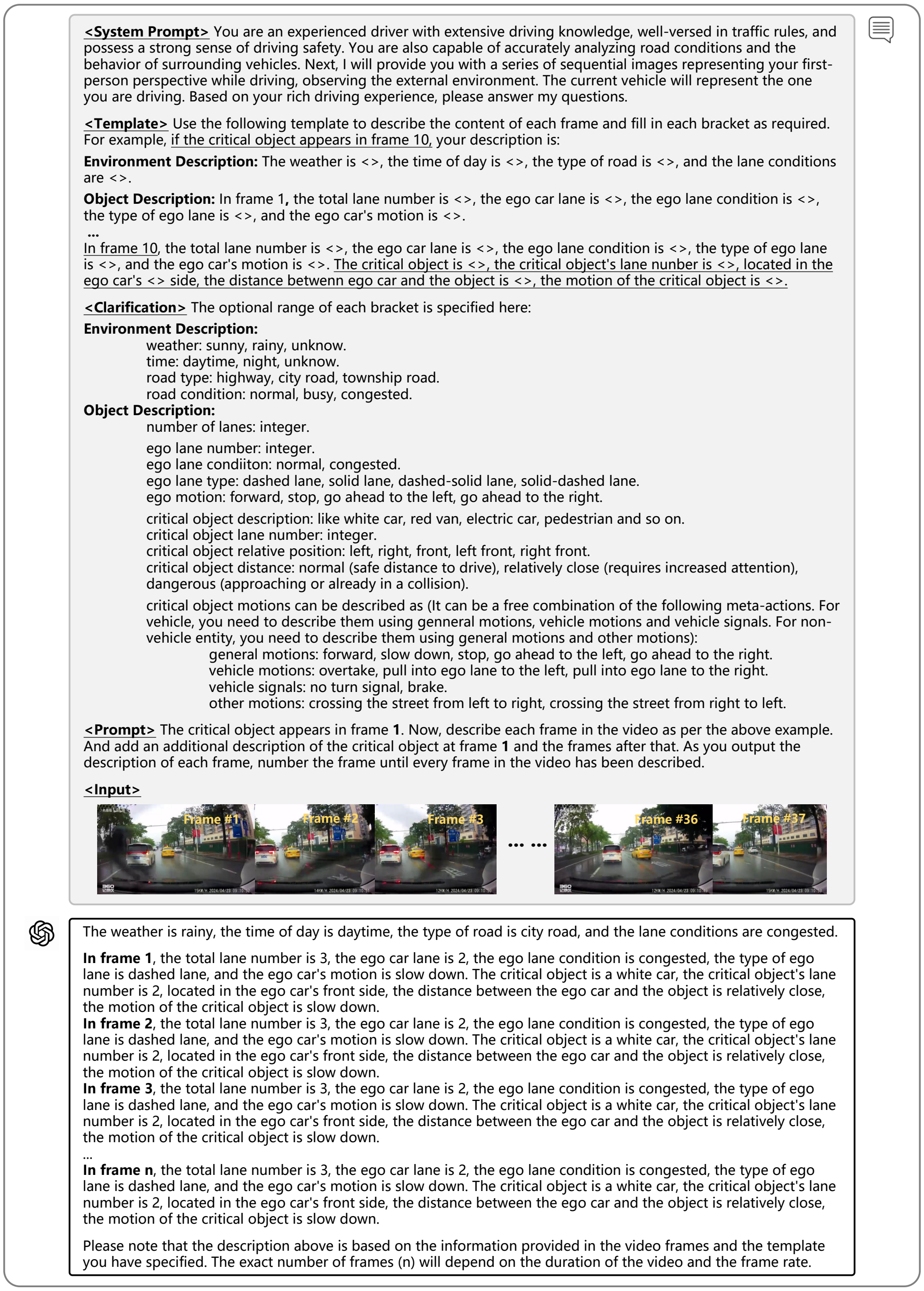
<System Prompt> You are an experienced driver with extensive driving knowledge, well-versed in traffic rules, and possess a strong sense of driving safety. You are also capable of accurately analyzing road conditions and the behavior of surrounding vehicles. Next, I will provide you with a series of sequential images representing your first-person perspective while driving, observing the external environment. The current vehicle will represent the one you are driving. Based on your rich driving experience, please answer my questions.

<Prompt> This set of images is sampled at a rate of 2 frames per second. Based on these images, please infer whether the following phenomena occur: 1. Dangerous driving behavior: Actions by other vehicles or pedestrians that force your vehicle to take measures to avoid collisions. 2. Aggressive driving behavior: Deliberate actions by other vehicles that interfere with the normal driving of your vehicle. 3. Road congestion: A high volume of traffic on the current road, causing your vehicle to move slowly. When answering, simply use 0 or 1 to indicate whether a phenomenon occurred. For example, if the scenario shown in the video includes dangerous driving behavior and road congestion but no aggressive driving behavior, your response would be 101. The first 1 indicates the occurrence of dangerous driving behavior, the second 0 indicates no aggressive driving behavior, and the third 1 indicates road congestion. Now, let's think step by step. For the input video, your output result is:

<Input>

101 **<Expected response: 110>**

Fig. 5. The experimental results for the main task are shown in the figure. We use video frames as input and ask the VLMs to identify dangerous driving, aggressive driving, and obstructive driving in the video. To simplify result analysis, VLMs are required to output a binary response (0 or 1).



<System Prompt> You are an experienced driver with extensive driving knowledge, well-versed in traffic rules, and possess a strong sense of driving safety. You are also capable of accurately analyzing road conditions and the behavior of surrounding vehicles. Next, I will provide you with a series of sequential images representing your first-person perspective while driving, observing the external environment. The current vehicle will represent the one you are driving. Based on your rich driving experience, please answer my questions.

<Template> Use the following template to describe the content of each frame and fill in each bracket as required. For example, if the critical object appears in frame 10, your description is:

Environment Description: The weather is <>, the time of day is <>, the type of road is <>, and the lane conditions are <>.

Object Description: In frame 1**,** the total lane number is <>, the ego car lane is <>, the ego lane condition is <>, the type of ego lane is <>, and the ego car's motion is <>.

...

In frame 10, the total lane number is <>, the ego car lane is <>, the ego lane condition is <>, the type of ego lane is <>, and the ego car's motion is <>. The critical object is <>, the critical object's lane nunber is <>, located in the ego car's <> side, the distance betwenn ego car and the object is <>, the motion of the critical object is <>.

<Clarification> The optional range of each bracket is specified here:

Environment Description:
- weather: sunny, rainy, unknow.
- time: daytime, night, unknow.
- road type: highway, city road, township road.
- road condition: normal, busy, congested.

Object Description:
- number of lanes: integer.
- ego lane number: integer.
- ego lane condiiton: normal, congested.
- ego lane type: dashed lane, solid lane, dashed-solid lane, solid-dashed lane.
- ego motion: forward, stop, go ahead to the left, go ahead to the right.
- critical object description: like white car, red van, electric car, pedestrian and so on.
- critical object lane number: integer.
- critical object relative position: left, right, front, left front, right front.
- critical object distance: normal (safe distance to drive), relatively close (requires increased attention), dangerous (approaching or already in a collision).
- critical object motions can be described as (It can be a free combination of the following meta-actions. For vehicle, you need to describe them using genneral motions, vehicle motions and vehicle signals. For non-vehicle entity, you need to describe them using general motions and other motions):
  - general motions: forward, slow down, stop, go ahead to the left, go ahead to the right.
  - vehicle motions: overtake, pull into ego lane to the left, pull into ego lane to the right.
  - vehicle signals: no turn signal, brake.
  - other motions: crossing the street from left to right, crossing the street from right to left.

<Prompt> The critical object appears in frame **1**. Now, describe each frame in the video as per the above example. And add an additional description of the critical object at frame **1** and the frames after that. As you output the description of each frame, number the frame until every frame in the video has been described.

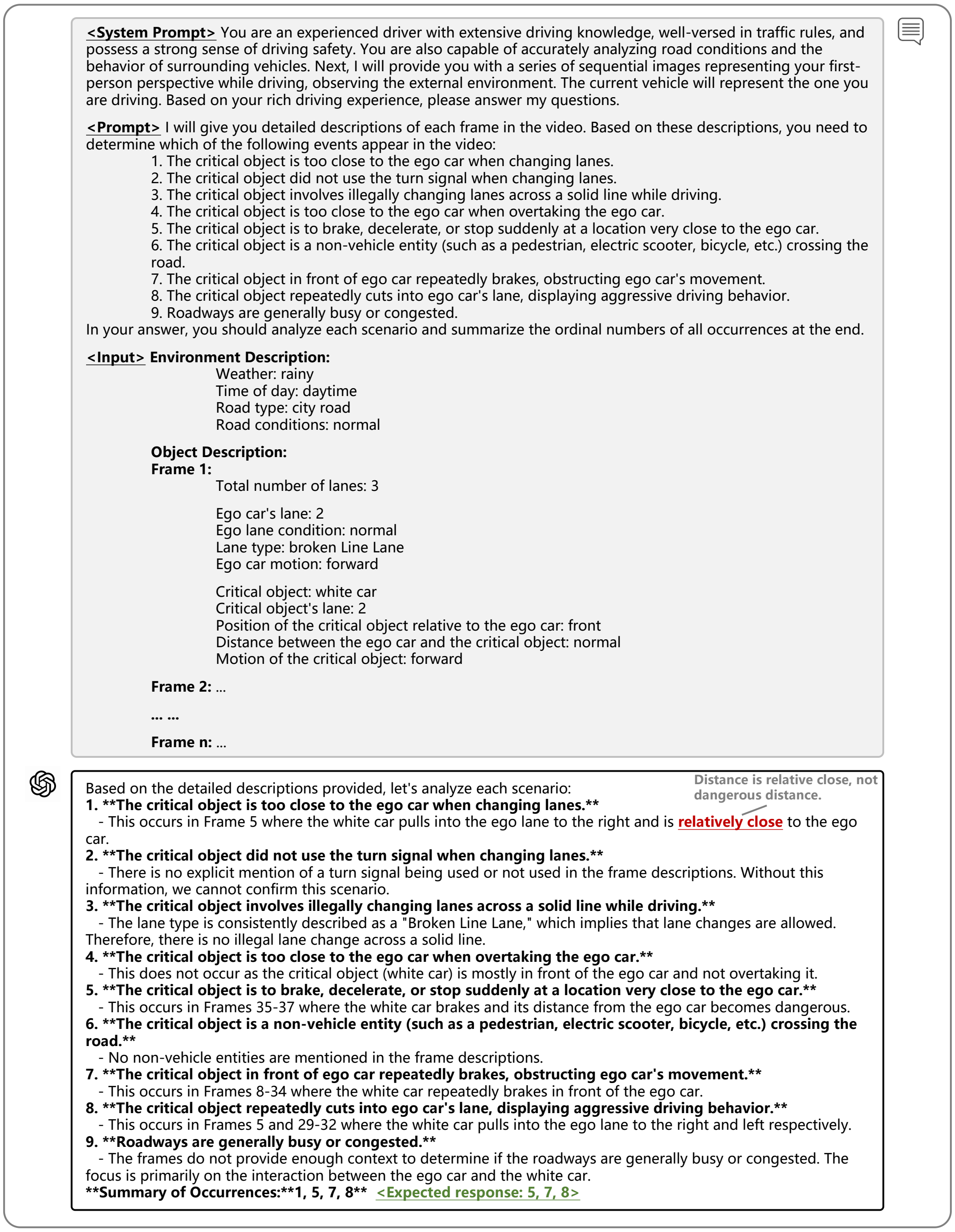
<Input>

The weather is rainy, the time of day is daytime, the type of road is city road, and the lane conditions are congested.

In frame 1, the total lane number is 3, the ego car lane is 2, the ego lane condition is congested, the type of ego lane is dashed lane, and the ego car's motion is slow down. The critical object is a white car, the critical object's lane number is 2, located in the ego car's front side, the distance between the ego car and the object is relatively close, the motion of the critical object is slow down.
In frame 2, the total lane number is 3, the ego car lane is 2, the ego lane condition is congested, the type of ego lane is dashed lane, and the ego car's motion is slow down. The critical object is a white car, the critical object's lane number is 2, located in the ego car's front side, the distance between the ego car and the object is relatively close, the motion of the critical object is slow down.
In frame 3, the total lane number is 3, the ego car lane is 2, the ego lane condition is congested, the type of ego lane is dashed lane, and the ego car's motion is slow down. The critical object is a white car, the critical object's lane number is 2, located in the ego car's front side, the distance between the ego car and the object is relatively close, the motion of the critical object is slow down.
...
In frame n, the total lane number is 3, the ego car lane is 2, the ego lane condition is congested, the type of ego lane is dashed lane, and the ego car's motion is slow down. The critical object is a white car, the critical object's lane number is 2, located in the ego car's front side, the distance between the ego car and the object is relatively close, the motion of the critical object is slow down.

Please note that the description above is based on the information provided in the video frames and the template you have specified. The exact number of frames (n) will depend on the duration of the video and the frame rate.

Fig. 6. The experimental results for sub-task 1 are shown in the figure. We provide a response template and ask VLMs to describe each frame of the input video. Since the critical object does not appear in every frame of the video, we inform VLMs in advance about which frames contain the critical object. VLMs are required to describe the critical object in these frames and the following ones. The results show that VLMs fail to provide complete responses, preventing us from performing quantitative analysis.



<System Prompt> You are an experienced driver with extensive driving knowledge, well-versed in traffic rules, and possess a strong sense of driving safety. You are also capable of accurately analyzing road conditions and the behavior of surrounding vehicles. Next, I will provide you with a series of sequential images representing your first-person perspective while driving, observing the external environment. The current vehicle will represent the one you are driving. Based on your rich driving experience, please answer my questions.

<Prompt> I will give you detailed descriptions of each frame in the video. Based on these descriptions, you need to determine which of the following events appear in the video:

1. The critical object is too close to the ego car when changing lanes.
2. The critical object did not use the turn signal when changing lanes.
3. The critical object involves illegally changing lanes across a solid line while driving.
4. The critical object is too close to the ego car when overtaking the ego car.
5. The critical object is to brake, decelerate, or stop suddenly at a location very close to the ego car.
6. The critical object is a non-vehicle entity (such as a pedestrian, electric scooter, bicycle, etc.) crossing the road.
7. The critical object in front of ego car repeatedly brakes, obstructing ego car's movement.
8. The critical object repeatedly cuts into ego car's lane, displaying aggressive driving behavior.
9. Roadways are generally busy or congested.

In your answer, you should analyze each scenario and summarize the ordinal numbers of all occurrences at the end.

<Input> **Environment Description:**

Weather: rainy
Time of day: daytime
Road type: city road
Road conditions: normal

Object Description:
Frame 1:

Total number of lanes: 3

Ego car's lane: 2
Ego lane condition: normal
Lane type: broken Line Lane
Ego car motion: forward

Critical object: white car
Critical object's lane: 2
Position of the critical object relative to the ego car: front
Distance between the ego car and the critical object: normal
Motion of the critical object: forward

Frame 2: ...

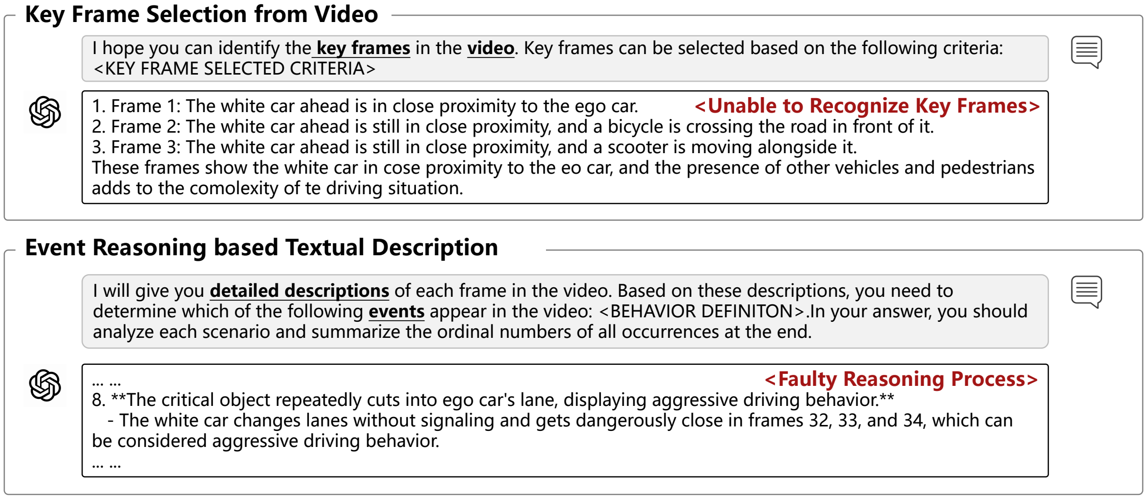
... ...

Frame n: ...

Distance is relative close, not dangerous distance.

Based on the detailed descriptions provided, let's analyze each scenario:
1. **The critical object is too close to the ego car when changing lanes.**
 - This occurs in Frame 5 where the white car pulls into the ego lane to the right and is relatively close to the ego car.
2. **The critical object did not use the turn signal when changing lanes.**
 - There is no explicit mention of a turn signal being used or not used in the frame descriptions. Without this information, we cannot confirm this scenario.
3. **The critical object involves illegally changing lanes across a solid line while driving.**
 - The lane type is consistently described as a "Broken Line Lane," which implies that lane changes are allowed. Therefore, there is no illegal lane change across a solid line.
4. **The critical object is too close to the ego car when overtaking the ego car.**
 - This does not occur as the critical object (white car) is mostly in front of the ego car and not overtaking it.
5. **The critical object is to brake, decelerate, or stop suddenly at a location very close to the ego car.**
 - This occurs in Frames 35-37 where the white car brakes and its distance from the ego car becomes dangerous.
6. **The critical object is a non-vehicle entity (such as a pedestrian, electric scooter, bicycle, etc.) crossing the road.**
 - No non-vehicle entities are mentioned in the frame descriptions.
7. **The critical object in front of ego car repeatedly brakes, obstructing ego car's movement.**
 - This occurs in Frames 8-34 where the white car repeatedly brakes in front of the ego car.
8. **The critical object repeatedly cuts into ego car's lane, displaying aggressive driving behavior.**
 - This occurs in Frames 5 and 29-32 where the white car pulls into the ego lane to the right and left respectively.
9. **Roadways are generally busy or congested.**
 - The frames do not provide enough context to determine if the roadways are generally busy or congested. The focus is primarily on the interaction between the ego car and the white car.
****Summary of Occurrences:**1, 5, 7, 8**** <Expected response: 5, 7, 8>

Fig. 7. The experimental results for sub-task 2 are shown in the figure. We convert the labels of the video into manual descriptions and use these as input for VLMs. VLMs are tasked with reasoning which behaviors appear in the video based on the descriptions (the collected videos contain 9 types of behaviors). The results show that VLMs can reason most behaviors correctly, but there are some errors in the details.



Key Frame Selection from Video

I hope you can identify the **key frames** in the **video**. Key frames can be selected based on the following criteria: <KEY FRAME SELECTED CRITERIA>

<Unable to Recognize Key Frames>

1. Frame 1: The white car ahead is in close proximity to the ego car.
2. Frame 2: The white car ahead is still in close proximity, and a bicycle is crossing the road in front of it.
3. Frame 3: The white car ahead is still in close proximity, and a scooter is moving alongside it.

These frames show the white car in cose proximity to the eo car, and the presence of other vehicles and pedestrians adds to the comolexity of te driving situation.

Event Reasoning based Textual Description

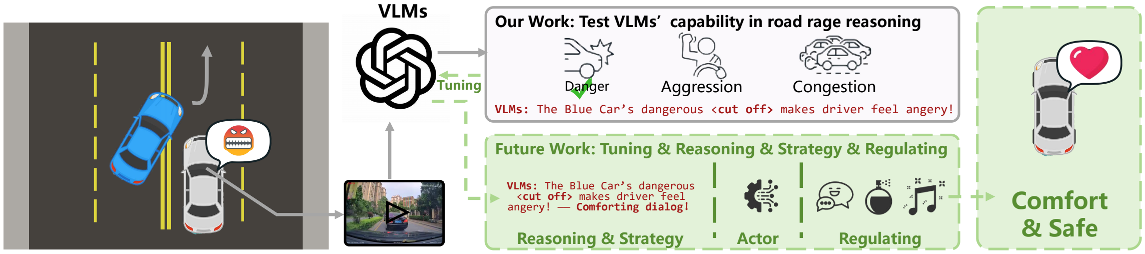
I will give you **detailed descriptions** of each frame in the video. Based on these descriptions, you need to determine which of the following **events** appear in the video: <BEHAVIOR DEFINITON>.In your answer, you should analyze each scenario and summarize the ordinal numbers of all occurrences at the end.

<Faulty Reasoning Process>

… …
8. **The critical object repeatedly cuts into ego car's lane, displaying aggressive driving behavior.**
   - The white car changes lanes without signaling and gets dangerously close in frames 32, 33, and 34, which can be considered aggressive driving behavior.

… …

Fig. 8. The results for selecting key frames from video input and performing event recognition based on textual description. The conclusion is that VLMs cannot identify key frames as requested, and they struggle with understanding spatial relationships in the textual modality to recognize road rage events.


VLMs
Tuning
Our Work: Test VLMs' capability in road rage reasoning
Danger
Aggression
Congestion
Future Work: Tuning & Reasoning & Strategy & Regulating
Reasoning & Strategy
Actor
Regulating
Comfort & Safe

VLMs: The Blue Car's dangerous <cut off> makes driver feel angery!

VLMs: The Blue Car's dangerous <cut off> makes driver feel angery! —— Comforting dialog!

Fig. 9. Future works