Responsibility Distribution Estimation in Ego-View Accident Videos with Multimodal Large Language Models

Anonymous ACL submission

Abstract

Recent studies on multimodal traffic accident understanding have mainly relied on infrastructure-camera footage, satellite imagery, or structured crash records. However, such data sources are costly to deploy and maintain at large scale, and they cannot objectively capture what the driver was actually able to observe before the accident. In contrast, ego-view accident videos directly represent the driver's visual perspective, making them suitable for reasoning about avoidability and driver responsibility. In this paper, we introduce *responsibility distribution estimation* for ego-view traffic accident videos, a new task in which a model predicts the percentage of responsibility assigned to each involved agent. We construct an LLM-assisted responsibility annotation pipeline and fine-tune multimodal large language models under multiple input settings, including raw frames, segmentation-enhanced input, and textual descriptions. Experimental results establish a strong initial benchmark, demonstrating that multimodal LLMs can effectively perform this nuanced, constraint-based reasoning task. Our findings suggest that ego-centric accident videos provide a promising foundation for socially and legally meaningful multimodal reasoning beyond conventional accident classification and explanation tasks.

1 Introduction

Recent advances in large language models (LLMs) and multimodal large language models (MLLMs) have enabled substantial progress in traffic accident analysis, including crash detection, severity prediction, and causal explanation. Existing approaches have primarily relied on structured crash reports, satellite imagery, or infrastructure-camera footage. While these settings are useful for scene-level monitoring, they suffer from two important limitations.

First, infrastructure-based systems are costly to deploy and maintain at scale. Traffic monitoring cameras and high-resolution sensing infrastructure are typically limited to specific intersections or urban regions, making comprehensive coverage difficult. Second, these external viewpoints cannot objectively represent what the driver was actually able to observe before the accident. As a result, they are less suitable for reasoning about avoidability and driver responsibility from the perspective of the vehicle involved in the accident.

In contrast, ego-view accident videos directly capture the driver's visual perspective. This perspective enables driver-centered reasoning, such as whether the ego driver had sufficient time to react, whether another agent suddenly entered the scene, and how responsibility should be distributed among the involved participants. Despite this unique property, prior work on ego-centric accident understanding has mainly focused on accident cause explanation and prevention-oriented reasoning, leaving responsibility allocation largely unexplored.

In this work, we propose *responsibility distribution estimation* for ego-view traffic accident videos. Given an accident video and a set of involved agents, the goal is to predict a responsibility distribution over the agents rather than a single accident label or cause. For example, instead of predicting that a pedestrian alone caused the accident, the model may estimate that the pedestrian is 60% responsible while the ego driver is 40% responsible. This formulation allows models to represent shared responsibility and more closely reflects real-world traffic reasoning. To study this problem, we construct an LLM-assisted responsibility labeling pipeline to generate proportional responsibility distributions. We establish the first benchmark for this task by evaluating multimodal models under different input conditions, including image-only input, segmentation-enhanced input, and textual descriptions.

Our work makes three contributions. First, we introduce responsibility distribution estimation as

a new task for multimodal accident understanding. Second, we highlight the importance of ego-centric accident videos as a foundation for driver-centered and responsibility-aware reasoning. We establish the first benchmark for this task by evaluating multimodal models under different input conditions.

2 Related Work

Recent work has explored the use of LLMs for traffic crash analysis. CrashLLM (Fan et al., 2024) formulates traffic crash records as language and investigates LLM-based prediction and causal analysis for accident outcomes. Other work has used LLMs with chain-of-thought prompting (Wei et al., 2022) and prompt engineering for crash severity prediction (Zhen et al., 2024), showing that language models can reason over environmental conditions, driver behavior, and vehicle-related factors. CrashSage (Zhen and Yang, 2025) further proposes an LLM-centered framework for contextual and interpretable crash analysis, combining data verbalization, augmentation, fine-tuning, and attribution-based explanation. These studies demonstrate the potential of LLMs for interpretable accident analysis, but they primarily focus on severity prediction, outcome classification, or causal factor identification rather than proportional responsibility allocation among agents.

Domain-adapted MLLMs have been applied to remote sensing imagery for interpretable road traffic accident analysis (He et al., 2025), where high-resolution overhead images are combined with structured accident data. Other work evaluates MLLMs for accident detection from infrastructure-camera imagery (Skender et al., 2025). These approaches are valuable for accident detection and scene-level explanation, but their visual perspectives differ from egocentric driving videos.

The MM-AU dataset introduced ego-view accident video understanding, emphasizing accident causes and possible prevention strategies from a driver-centered perspective (Fang et al., 2024). This setting is particularly relevant to responsibility reasoning because the video captures what the ego driver could observe before the accident. Prior use of MM-AU and related datasets has mainly focused on explaining accident causes or generating prevention-oriented descriptions. We extend this direction by asking models to estimate responsibility distributions over involved agents, moving from causal explanation to socially meaningful allocation of responsibility.

CADD (Shen and Wu, 2025) bridges behavioral understanding in dashcam videos with statute-based liability attribution by utilizing a framework to map driving actions directly to specific traffic laws for definitive liability judgments. While their work emphasizes strict legal and statutory grounding for liability, our approach focuses on estimating the continuous proportional responsibility distribution among involved agents using generative multimodal reasoning.

Compared with prior work on crash severity prediction, accident detection, and natural-language causal explanation, our work is significantly different in following ways. First, we focus on ego-centric accident videos rather than structured crash records, satellite imagery, or infrastructure-camera images. Second, we predict a responsibility distribution rather than a single accident type, severity label, or textual cause. Although our current labels are LLM-assisted and should not be treated as legal ground truth, the task provides a practical framework for studying whether multimodal models can produce structured, constraint-satisfying responsibility judgments from accident videos.

3 Method

3.1 Task Definition

We propose *Responsibility Distribution Estimation* for ego-view traffic accident videos. Given an accident video V and a set of involved agents $A = \{a_1, a_2, \ldots, a_n\}$, the goal is to predict a responsibility distribution $\mathbf{r} = [r_1, r_2, \ldots, r_n]$, where each $r_i \in [0, 100]$ represents the percentage of responsibility assigned to agent a_i, subject to $\sum_{i=1}^{n} r_i = 100$.

Unlike conventional accident classification tasks, our objective is not to identify a single cause or responsible agent, but rather to estimate how responsibility is distributed among multiple participants. In our setting, the agents may include the ego vehicle, other vehicles, pedestrians, bicycles, or motorcycles. Because the videos are ego-centric, the task additionally requires reasoning about what information was available to the ego driver before the accident. We also require the model to identify which party is primarily responsible and to label the type of the other party.

3.2 Dataset Construction

We build our dataset on top of MM-AU (Fang et al., 2024), which provides ego-view dashcam accident videos with text annotations describing the accident scene, its causes, and suggested preventive measures. Since MM-AU does not include responsibility labels, we construct an automated annotation pipeline with LLM as an annotator.

For each accident video, we uniformly sample $k = 8$ frames spanning the annotated abnormal event window, ensuring the accident frame is always included. We present these frames to LLM alongside the scene metadata and all available text annotations. The model is prompted to output (i) the integer percentages of responsibility assigned to the ego vehicle and other parties, (ii) which party is the primary responsible party, (iii) the category of the other parties (`pedestrian`, `other_vehicle`, `motorcycle_cyclist`, `environment`, or `unknown`), and (iv) a rationale grounding the judgment in the visual and textual evidence. We use temperature zero to maximize label consistency, and ensure that percentages sum to exactly 100. The resulting dataset is roughly balanced between ego-primary (52%) and other-party-primary (48%) cases. The most frequent other-party types are other vehicles (58%) and motorcycle cyclists (33%), with pedestrians accounting for 7%.

3.3 Model and Training

We fine-tune Qwen3-VL-4B (Team, 2025) on the annotated responsibility dataset with LoRA adaptation (Hu et al., 2022) to all linear projection layers with rank $r = 16$, scaling factor $\alpha = 32$, and dropout 0.05. Given a multimodal input X consisting of k accident frames and an instruction prompt, the model autoregressively generates a JSON-formatted output containing the predicted responsibility percentages, the primary responsible party, the other-party label, and a brief rationale.

Each training example follows an instruction-tuning format. The user turn contains the case metadata, text annotations, and $k = 10$ sampled frames. The assistant turn contains the target output derived from the LLM annotations. Distribution validity is enforced implicitly: the prompt instructs the model that the percentages must sum up to 100, and all training data satisfy this constraint, so the model learns the requirement from the data.

Model	Input	EM	W-10	W-20	MAE	PP
GPT-5 (ZS)	RGB	38.5	79.1	86.8	12.1	89.0
GPT-5 (ZS)	RGB+Seg	28.6	79.1	87.9	12.7	89.0
GPT-5 (ZS)	Text	23.1	45.1	57.1	24.9	62.6
Qwen (FT)	RGB	**79.1**	**90.1**	**95.6**	**4.9**	**96.7**
Qwen (FT)	RGB+Seg	**79.1**	**90.1**	94.5	5.5	95.6
Qwen (FT)	Text	24.2	51.6	57.1	35.1	59.3

Table 1: Test-set results. EM: Exact Match; W-10/W-20: within 10/20 percentage points; MAE: mean absolute error on ego-vehicle percentage; PP: Primary Party accuracy. ZS: zero-shot; FT: LoRA fine-tuned.

4 Experiment

4.1 Setting

Input: To analyze which sources of information contribute to responsibility estimation, we evaluate three input conditions. **RGB** provides the model with $k = 10$ uniformly sampled accident frames together with all text annotations. **RGB+Seg** replaces the RGB frames with semantic segmentation overlays extracted from MM-AU while keeping the same text annotations, allowing us to test whether structured visual representations are more informative than raw appearance. **Text-only** provides only the scene description and no images, isolating the contribution of visual input.

Baseline: We compare our fine-tuned Qwen3-VL-4B against GPT-5 used in a zero-shot setting under the same three input conditions. For each condition, the same structured JSON prompt is used for both models, ensuring a fair comparison. The zero-shot baseline assesses the out-of-the-box capability of a strong frontier model on our task.

Evaluation metrics: We report the following metrics. **Exact Match** (%): the proportion of cases where the predicted ego-vehicle responsibility percentage equals the ground-truth label exactly. **Within-10%** and **Within-20%**: the proportion of cases where the absolute error is at most 10 or 20 percentage points, measuring practical tolerance. **MAE** and **RMSE**: mean absolute error and root mean squared error of the ego-vehicle responsibility percentage. **Primary Party Acc.** (%): accuracy of predicting which party bears primary responsibility, a classification sub-task within the distribution estimation problem.

4.2 Results and Discussion

Table 1 presents the main results. Our fine-tuned Qwen3-VL-4B demonstrates that models can successfully learn to allocate proportional responsibility, successfully adapting to the constraints of the

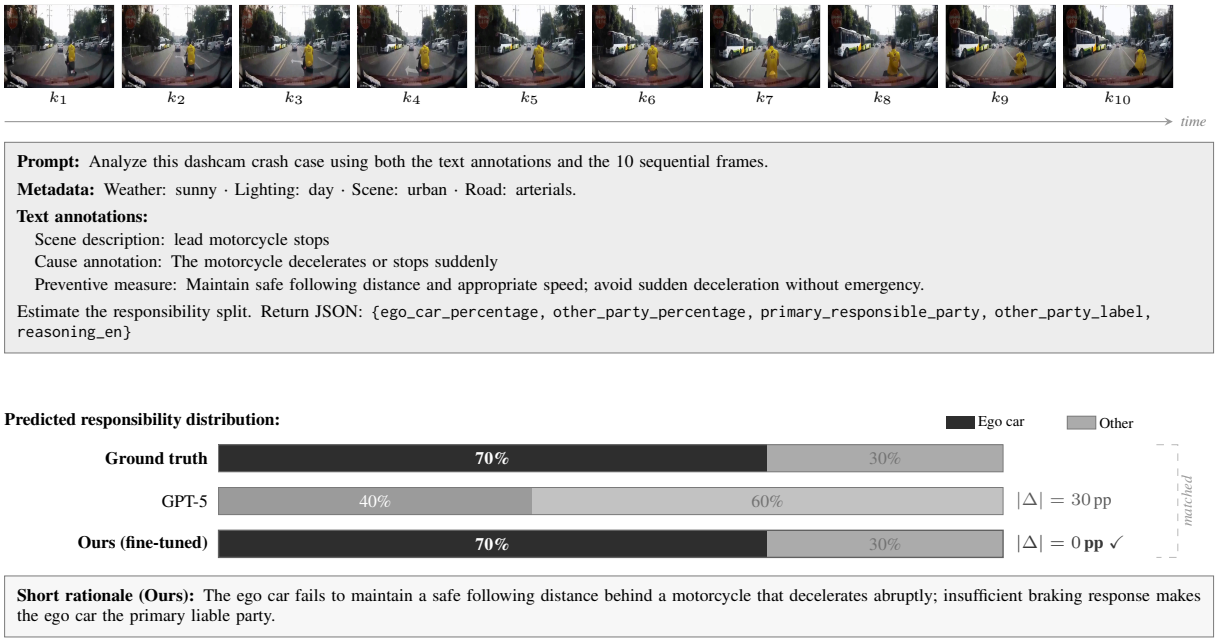

Figure 1: Qualitative input–output comparison on a held-out test case from MM-AU. **Top**: ten uniformly sampled ego-view frames spanning the accident window (k_1–k_{10}). **Middle**: the actual prompt given to both models, including scene metadata and text annotations. **Bottom**: predicted responsibility distributions (*dark* = ego car; *light* = other party) for GPT-5 and our fine-tuned model, compared with the ground truth. GPT-5 misidentifies the primary responsible party and incurs a 30-percentage-point error; our model matches the ground truth exactly ($|\Delta| = 0$ pp).

task compared to frontier models. Under RGB input, the fine-tuned model achieves an Exact Match of 79.1% and a Primary Party accuracy of 96.7%, compared to 38.5% and 89.0% for GPT-5 zero-shot. MAE drops from 12.1 to 4.9 percentage points, a reduction of more than half.

As expected, both the fine-tuned model and GPT-5 degrade sharply when visual input is removed. Under text-only input, GPT-5 MAE more than doubles (12.1 $\rightarrow$ 24.9) and Primary Party accuracy drops to 62.6%. The fine-tuned model degrades even more severely on MAE (4.9 $\rightarrow$ 35.1). These results confirm that ego-view video frames carry information about responsibility that cannot be recovered from textual descriptions alone.

Figure 1 shows a qualitative input-output example. Both GPT-5 and our fine-tuned model receive the same ego-view frames, metadata, text annotations, and structured prompt. While GPT-5 reverses the primary responsible party, our model predicts the same responsibility distribution as the ground-truth label, demonstrating that training with LLM-annotated allocation helps the model better map multimodal accident evidence to the responsibility distribution. Semantic segmentation overlays yield comparable or slightly worse results for both models, with Primary Party accuracy dropping by 1–2 points and MAE increasing slightly, suggesting that the segmentation representation hardly adds information beyond what RGB frames. Structured visual representations combined with more explicit reasoning scaffolds may be beneficial.

GPT-5 zero-shot already achieves high Within-10% (79.1%) under RGB input, indicating that the zero-shot model can reason roughly correctly about which direction responsibility falls. However, our fine-tuned model's Exact Match of 79.1% shows that fine-tuning on LLM-annotated labels substantially anchors the model to the specific percentage values present in the training distribution, rather than merely ranking parties correctly.

5 Conclusion & Future Work

We proposed a novel responsibility distribution estimation task for ego-view traffic accident videos to assign proportional responsibility to the involved agents. We constructed an LLM-assisted annotation and training pipeline, and our experiments demonstrate that multimodal models can successfully navigate the nuanced reasoning required for proportional responsibility allocation. Future research should further validate the LLM-assisted labels with human experts, expand the dataset to more diverse traffic scenarios, and incorporate legal context into responsibility estimation.

Limitations

This study has several limitations. First, the responsibility labels are generated with LLM assistance and should not be treated as human expert annotations or legal ground truth. The current annotation process does not explicitly account for jurisdiction-specific legal context, which limits its ability to distinguish causal contribution from legal responsibility.

Second, our model relies on sampled frames and available text annotations rather than explicit physical evidence such as speed, braking behavior, agent trajectories, or reaction time. The model may therefore miss fine-grained temporal cues that are important for responsibility estimation.

Third, our current formulation mainly considers responsibility allocation between the ego vehicle and another involved party. Accidents involving three or more interacting agents may require more complex responsibility decomposition, since the model must distinguish each agent's contribution and their interactions before the collision.

Ethical Considerations

The estimation of responsibility in traffic accidents is a highly sensitive task with significant legal, financial, and personal implications. While our proposed framework demonstrates the potential for driver-centered multimodal accident reasoning, it is crucial to emphasize that this system is an exploratory research prototype and must not be used for automated legal, judicial, or insurance decision-making. Primarily, our training dataset utilizes LLM-assisted responsibility labels; these serve as a proxy to evaluate multimodal constraint-satisfaction and reasoning, but they are not verified legal ground truths or human expert judgments. Furthermore, the model evaluates scenarios based purely on visual and text-annotated sequences, lacking the capacity to account for jurisdiction-specific statutory traffic laws or fine-grained physical telemetry, such as exact vehicle speed and braking force. Relying strictly on such a model in real-world settings risks automation bias, where a system's misinterpretation of an ego-view perspective could lead to an unjust distribution of fault. Therefore, any future real-world application of this technology necessitates a strict human-in-the-loop architecture, wherein the model functions solely as an assistive tool to help investigators triage multimodal evidence, rather than replacing the legally binding and nuanced judgments of traffic authorities. Finally, because dashcam videos inherently capture personally identifiable information such as license plates and faces, researchers expanding on this dataset must employ rigorous anonymization protocols to protect the privacy of all recorded agents.

References

Zhiwen Fan, Pu Wang, Yang Zhao, Yibo Zhao, B. Ivanovic, Zhangyang Wang, Marco Pavone, and Hao Frank Yang. 2024. Learning traffic crashes as language: Datasets, benchmarks, and what-if causal analyses. *ArXiv*, abs/2406.10789.

Jianwu Fang, Lei lei Li, Junfei Zhou, Junbin Xiao, Hongkai Yu, Chen Lv, Jianru Xue, and Tat-Seng Chua. 2024. Abductive ego-view accident video understanding for safe driving perception. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22030–22040.

Bing He, Wei He, Qing Chang, Wen Luo, and Ling Xiao. 2025. Domain-adapted mllms for interpretable road traffic accident analysis using remote sensing imagery. *ISPRS Int. J. Geo Inf.*, 15:8.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Yunfei Shen and Zhongcheng Wu. 2025. Cadd: A chinese traffic accident dataset for statute-based liability attribution. *ArXiv*, abs/2511.11715.

Ilhan Skender, Kailin Tong, Selim Solmaz, and Daniel Watzenig. 2025. Investigating traffic accident detection by using multimodal large language models. *2025 IEEE International Automated Vehicle Validation Conference (IAVVC)*, pages 1–7.

Qwen Team. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc.

Hao Zhen, Yucheng Shi, Yongcan Huang, Jidong J. Yang, and Ninghao Liu. 2024. Leveraging large language models with chain-of-thought and prompt engineering for traffic crash severity analysis and inference. *ArXiv*, abs/2408.04652.

Hao Zhen and Jidong J. Yang. 2025. Crashsage: A large language model-centered framework for contextual and interpretable traffic crash analysis. *ArXiv*, abs/2505.07853.