Understanding Real-World Traffic Safety through RoadSafe365 Benchmark

Xinyu Liu[1], Darryl C. Jacob[1], Yuxin Liu[2], Xinsong Du[3], Muchao Ye[4], Bolei Zhou[2], Pan He[1*]
[1]Auburn University [2]University of California, Los Angeles
[3]Harvard Medical School [4]The University of Iowa

Abstract

Although recent traffic benchmarks have advanced multimodal data analysis, they generally lack systematic evaluation aligned with official safety standards. To fill this gap, we introduce RoadSafe365, a large-scale vision-language benchmark that supports fine-grained analysis of traffic safety from extensive and diverse real-world video data collections. Unlike prior works that focus primarily on coarse accident identification, RoadSafe365 is independently curated and systematically organized using a hierarchical taxonomy that refines and extends foundational definitions of crash, incident, and violation to bridge official traffic safety standards with data-driven traffic understanding systems. RoadSafe365 provides rich attribute annotations across diverse traffic event types, environmental contexts, and interaction scenarios, yielding 36,196 annotated clips from both dashcam and surveillance cameras. Each clip is paired with multiple-choice question–answer sets, comprising 864K candidate options, 8.4K unique answers, and 36K detailed scene descriptions that collectively designed for vision–language understanding and reasoning. We establish strong baselines and observe consistent gains when fine-tuning on RoadSafe365. Cross-domain experiments on both real and synthetic datasets further validate its effectiveness. Designed for large-scale training and standardized evaluation, RoadSafe365 provides a comprehensive benchmark to advance reproducible research in real-world traffic safety analysis.

1. Introduction

Threats to traffic safety, such as crashes, incidents, and violations, remain a pressing issue, resulting in substantial losses of lives, property, and societal welfare. In 2024 alone, motor vehicle crashes in the United States led to an estimated 39,345 deaths and nearly 2.2 million injuries [29], underscoring the urgent need for effective prevention to enhance public safety and traffic management [4, 12, 28, 54]. Although autonomous driving has advanced rapidly, models trained on routine driving datasets, despite their strong performance in general traffic understanding, struggle to recognize critical edge cases and rare failure scenarios such as traffic accidents and violations. A deeper understanding of traffic is crucial for anticipating risks and improving road safety. Encouragingly, recent advances in Multimodal Large Language Models (MLLMs), especially Vision-Language Models (VLMs), show strong ability to interpret complex scenes and generate detailed, context-aware explanations, outperforming conventional perception models that rely on black-box visual features [13, 15, 17, 37, 38, 45] for the general-purpose traffic understanding.

However, a major limitation remains: there is an absence of large-scale, high-quality datasets for training and evaluating vision–language models in traffic scenarios that require domain-specific, fine-grained safety understanding. Although several traffic accident datasets have been introduced to address this gap, they remain limited in scope: they often focus on specific aspects such as vulnerable road users (VRUs), perception, planning, or model robustness, but they lack systematically annotated datasets that support detailed accident causality reasoning and safety-critical evaluation of VLMs [19, 24, 31, 42, 43, 46]. More critically, existing datasets seldom organize traffic accidents and violations in a manner consistent with official safety standards, e.g., the definitions outlined in the Manual on Classification of Motor Vehicle Traffic Crashes (ANSI D16) by the National Safety Council (NSC) [30]. Without such datasets, the gap between general-purpose and safety-critical tasks can lead to inaccurate or inconsistent responses, as VLMs, even advanced ones such as GPT-4o [14], Qwen2-VL [41], InternVL [8], and Gemini-1.5-Pro [40], may misinterpret the visual content of traffic accidents. This is because accidents and violations stem from long-tailed, rare-event distributions that differ markedly from everyday traffic patterns.

To address these gaps, we introduce RoadSafe365, a large-scale vision–language benchmark for fine-grained traffic safety understanding built upon extensive real-world video data collected from public platforms across multi-

*Corresponding author. Email: pan.he@auburn.edu.

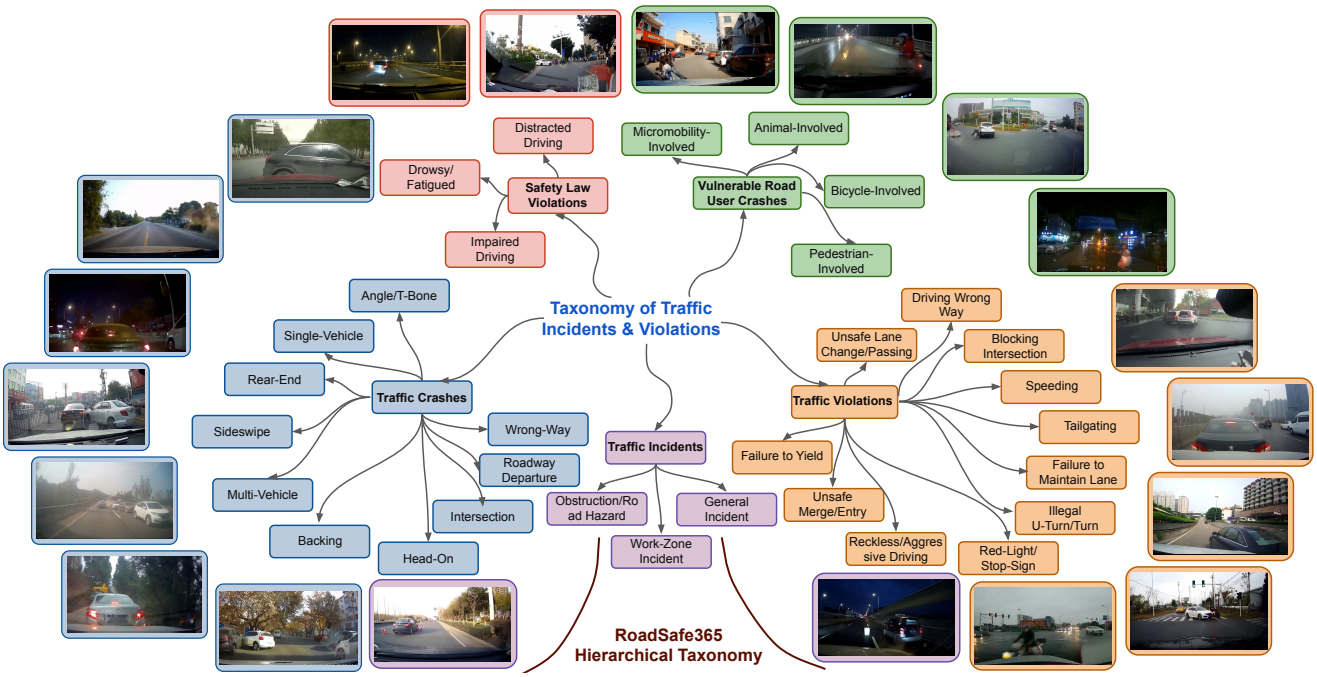

Figure 1. Overview of the taxonomy for traffic safety understanding defined in the RoadSafe365 benchmark.

ple countries and regions, forming a diverse and high-quality visual foundation. We propose a new hierarchical taxonomy (as shown in Fig. 1) aligned with official safety standards and develop a rigorous annotation pipeline that produces comprehensive, human-verified annotations to support detailed reasoning tasks. This enables RoadSafe365 to offer rich attribute annotations containing diverse event types, environmental conditions, and interaction scenarios, with 36,196 annotated clips captured from both dashcam and surveillance perspectives. Each clip is accompanied by multiple-choice Visual Question Answering (VQA) sets covering five safety-critical categories, comprising 864K candidate options, 8.4K unique answers, and 36K detailed scene descriptions, collectively enabling comprehensive vision–language understanding and reasoning. As a foundational benchmark for training and evaluating MLLMs, RoadSafe365 aims to advance fine-grained accident understanding and generalizability in safety-critical scenarios. Our main contributions are summarized as the following:

- We introduce RoadSafe365, a large-scale vision-language benchmark for traffic safety understanding that includes 36,196 independently collected videos from both dashcam and surveillance perspectives.
- We propose a hierarchical taxonomy that refines and extends foundational traffic safety concepts of *crash*, *incident*, and *violation* to align official safety standards with data-driven fine-grained traffic understanding.
- Leveraging the new taxonomy in RoadSafe365, we design a structured annotation pipeline that provides detailed QA attributes and narrative captions aligned with traffic safety standards. Comprehensive evaluations of state-of-the-art VLMs show that fine-tuning on RoadSafe365 improves both VQA and captioning performance while generalizing to unseen accident datasets, including a public real-world benchmark and another new synthetic dataset RoadSafe365-Synthetic. We aim to promote reproducible and systematic research in traffic safety.

2. Related Work

Accident and Driving Scene Understanding Datasets. Recent works have introduced various datasets to support research on accident and driving understanding, which have incorporated language to support complex reasoning. We find that there is a lack of datasets in accident scene understanding. To illustrate, MAPLM [7] provides a large-scale benchmark combining 2D images, 3D LiDAR, and HD map annotations focusing on map and scene-level reasoning, but ignores accident or violation analysis; MetaVQA [43] builds an embodied VQA benchmark for spatial and temporal reasoning in dynamic driving scenes, but it does not focus on detailed accident causality or standard safety-based categories; and NuPlanQA [34] explores multi-view and multimodal understanding in driving environments, but it mainly targets general perception instead of structured accident and violation reasoning. Although datasets such as

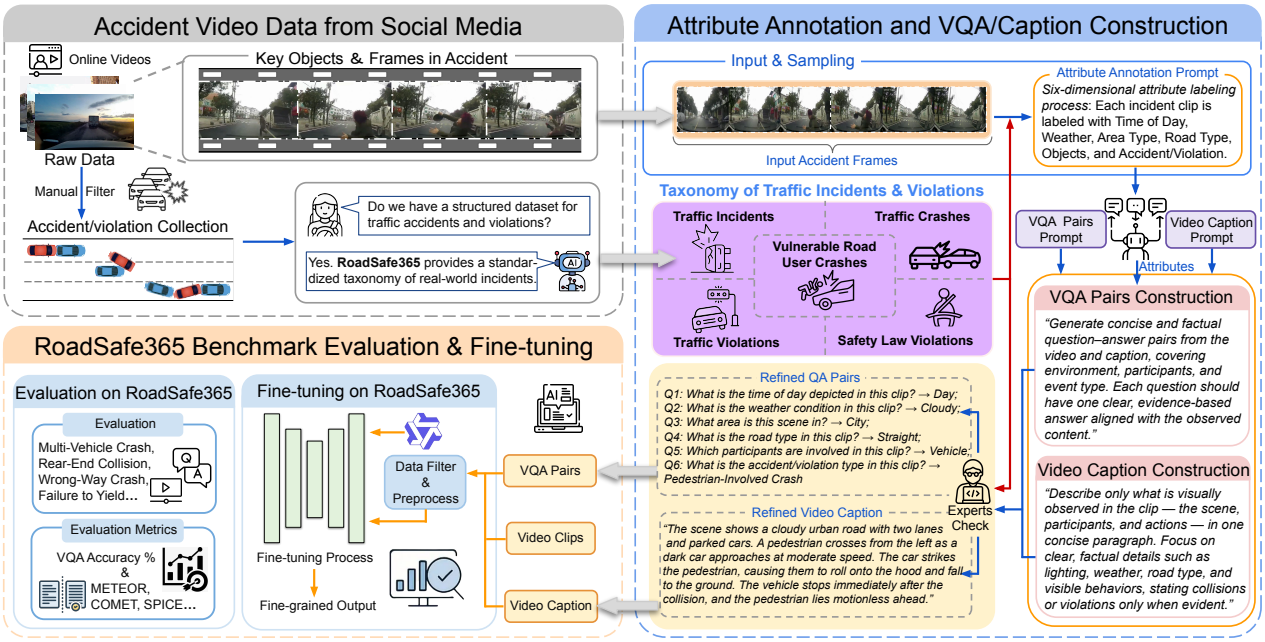

Figure 2. Overview of RoadSafe365: Collection, Annotation, Training, and Evaluation

DADA-2000 [10], MM-AU [5], and TAU-106K [56] offer real-world videos with causal annotations for accident scene understanding, they primarily aggregate publicly available videos with coarse labeling, limiting systematic analysis aligned with official safety standards. Also, datasets such as VRU-Accident [19] and RoadSocial [33] target specific scenarios, such as VRUs or road events from social media, and lack comprehensive coverage of diverse accidents. To overcome these limitations, we introduce RoadSafe365, a large-scale dataset offering standardized and comprehensive annotations for diverse traffic accidents, enabling fine-grained safety analysis aligned with official traffic safety standards.

Multimodal Large Language Models. Traffic research, particularly in autonomous driving, has increasingly embraced MLLMs that integrate visual perception into LLMs, enabling textual reasoning over visual scenes. Example MLLMs applied to driving tasks include foundation models like InternVL [8], Qwen-VL [41], and LLaVA [20], and specialized models such as DriveVLM [39] and LMDrive [36]. More recent frameworks include VLP [31], a vision-language planning framework that uses language models to improve trajectory planning in autonomous vehicles, and DriveGPT4-V2 [49], a closed-loop driving pipeline integrating LLMs to enable language-guided decision-making and multi-stage reasoning. Despite these advances, most works focus on normal driving scenes, with training data lacking diverse and realistic accident scenarios due to long-tailed, rare-event distributions—limiting their ability to reason about collisions, violations, and causal relationships. To bridge this gap, we introduce our dataset to provide the foundation for developing more reliable and robust MLLMs in safety-critical environments.

Reliable and Interpretable Accident Understanding. Given the importance of reliability and interpretability for MLLMs in traffic understanding, researchers have developed specialized assessment frameworks to evaluate these aspects. To illustrate, DriveBench [46] and AUTOTrust [47] provide systematic frameworks to assess model trustworthiness and robustness under corruptions, measuring dimensions such as safety, uncertainty, and fairness; AutoDrive-QA [18] introduces a multiple-choice QA benchmark focused on driving scenarios, aiming to improve standardized evaluation of vision-language models; and DriveArena [50] proposes a closed-loop generative simulation platform that evaluates reasoning and intervention performance of autonomous driving models. However, most existing studies emphasize benchmarking over large-scale, real-world data collection. While they provide valuable evaluation tools, they lack the detailed accident annotations essential for developing safety-critical models. Our RoadSafe365 fills this gap by serving as both a large-scale training source and a comprehensive benchmark, supporting the development of more reliable and interpretable models.

3. RoadSafe365

3.1. Video Data Collection and Preprocessing

Understanding traffic safety is vital for enhancing public safety and effective traffic management. However, existing public datasets are often limited in scale, diversity, and

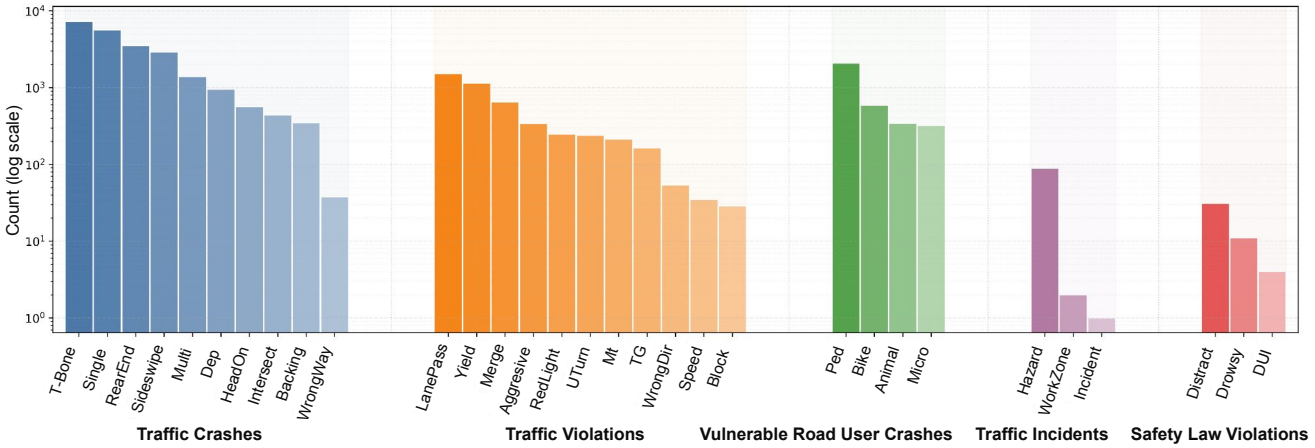

Figure 3. RoadSafe365 Annotation Taxonomy and Data Distribution. (a) shows the distribution (log scale) of the five Level-1 categories in our taxonomy. (b) illustrates the Level-2 subcategories nested within each Level-1 category.

visual quality. As Fig. 2 shows, to build a more comprehensive and realistic collection, we curated accident videos from Bilibili and X (formerly Twitter), two popular social platforms that host extensive dashcam and surveillance footage. Our data collection strategy entails identifying blogger channels focused on traffic scene understanding—many of which host extensive crash and violation footage—and complementing this with keyword-based searches to retrieve additional accident-related videos. This process facilitates large-scale acquisition of real-world traffic footage covering a wide range of weather, lighting, and traffic conditions. For each raw footage, we further segmented it into multiple short clips using CapCut [6], an efficient video editing tool, followed by manual verification to ensure that each segment clearly depicts a traffic accident or other safety-critical event. Irrelevant, duplicate, and low-quality clips, e.g., those with severe occlusion or unstable viewpoints, were removed to maintain high data quality and consistency. In total, we collected 36,196 accident clips, each lasting around 10–20 seconds, spanning urban intersections, highways, residential areas, and rural roads.

3.2. Taxonomy for Traffic Safety Understanding

As discussed, prior works primarily focus on coarse identification of traffic accidents or violations, lacking alignment with official safety standards established by traffic authorities. In this paper, as shown in Fig. 1, RoadSafe365 introduces a hierarchical taxonomy inspired by and adapted from widely recognized traffic incident standards, following definitions established by the NSC [30] to promote uniformity and comparability in motor vehicle traffic accident statistics. Specifically, we organize each clip by its event type following a two-level hierarchical structure:

Firstly, as illustrated in Fig. 3(a), we categorize all traffic events into five primary Level-1 classes: traffic crashes, traffic violations, VRU crashes, traffic incidents, and safety law violations. The rationale behind the Level-1 category design is to assess each event's impact based on person type, injury severity, and property damage. This structure enables detailed reasoning about questions such as whether a person was involved in a crash inside or outside a vehicle, the extent of property damage, the number of vehicles in transport involved, and the role of each person at the time of the crash, which provides more insights beyond existing benchmarks.

Secondly, each Level-1 category is further divided into multiple fine-grained Level-2 subcategories, as visualized in Fig. 3(b). For instance, *Traffic Crashes* include collision types such as rear-end, angle/T-bone (T-Bone), sideswipe, single-vehicle (Single), and intersection-related crashes. *Traffic Violations* capture risky driving behaviors such as unsafe lane changes or passing (LanePass), failure to yield (Yield), red-light violations (RedLight), aggressive driving, and tailgating (TG). *VRU Crashes* involve incidents with pedestrians (Ped), cyclists (Bike), animals, and micromobility (Micro) users, while *Traffic Incidents* cover non-collision events such as roadway obstructions (Hazard) and work-zone disruptions. Lastly, *Safety Law Violations* consider human-factor offenses such as distracted, drowsy, or impaired (DUI) driving.

Grounded in official safety standards, our hierarchical taxonomy enables structured accident analysis and causal reasoning, while existing benchmarks only use coarse, unsystematic event categories [19, 56].

3.3. Annotation Pipeline

To enable fine-grained analysis, we followed the proposed hierarchical taxonomy and developed an annotation pipeline that integrates structured attributes, question–answer pairs, and descriptive captions, as shown in Fig. 2. Each video is first converted into a short montage

of key frames and then annotated to ensure both structural consistency and rich contextual details.

Attribute Annotation. We extract six core attributes using Gemini-2.5-Pro [9] and manually verify each for accuracy and consistency. These attributes include time of day, weather, area type, road type, objects involved, and accident/violation category. The accident/violation label follows our two-level hierarchical structure consisting of a coarse class (e.g., collision, violation, or unknown) and a fine class aligned with standardized traffic categories (e.g., rear-end collision, unsafe lane change, failure to yield). When sufficient evidence is unavailable, annotators assign the label unknown rather than inferring intent, ensuring objectivity and reliability across all data sources.

Descriptive Captions. Each clip is also accompanied by a detailed textual description generated by GPT-4o [14], focusing exclusively on visually observable elements such as the scene context, participants, their actions and interactions, and the resulting outcomes. All generated captions were subsequently reviewed and refined by human annotators to ensure accuracy and clarity.

VQA Construction. To support question–answering tasks, the verified attributes are transformed into multiple-choice questions. Each question includes one correct answer and three distractors sampled from semantically similar categories to reduce trivial guessing. For multi-label questions (e.g., objects involved), distractors are generated by slightly modifying the correct set (e.g., drop-one or add-one). For accident or violation questions, we employ a curated combination bank to ensure that all answer choices are plausible and domain-relevant. Both automated validation and manual review are applied to maintain consistency, accuracy, and overall annotation quality.

All prompt templates and implementation details for attribute generation, captioning, and VQA construction are provided in the Appendix for interested readers.

3.4. Summary of RoadSafe365 Statistics

The RoadSafe365 dataset comprises 36,196 real-world traffic accident video clips, each annotated with structured attributes, multiple-choice question–answer pairs, and detailed descriptive captions. In total, it includes over 200K attribute labels, 36K dense captions, and more than 200K VQA pairs. The annotations offer diverse accident scenarios with fine-grained labels and detailed event descriptions that capture both spatial and temporal context across a range of scenarios.

4. Experiments

In this section, we evaluate several state-of-the-art MLLMs on the proposed RoadSafe365 benchmark across four core tasks: VQA, Dense Captioning, Fine-tuning Evaluation, and Cross-Domain Generalization, to analyze their performance and generalization capabilities.

4.1. Experimental Setup

We evaluate both open-source and closed-source MLLMs on the RoadSafe365 benchmark. These models include Qwen2.5-VL-7B [2], InternVL3-8B [8], LLaVA-NeXT-Video-7B [21], Gemini 2.5 Pro [9], and GPT-4o [14], covering a diverse range of model sizes and architectures. For fine-tuning, we adopt VAU-R1 [57] as the base model, leveraging its reinforcement learning–based optimization framework for video anomaly understanding.

4.2. Tasks and Evaluation Protocol

VQA Task. Each video is accompanied by multiple QA items spanning six categories: Time, Weather, Area, Road Type, Object, and Accident. Each question includes four candidate answers, one correct and three distractors. Models are prompted to select the correct option, and accuracy is computed for each category and averaged across all categories. This task emphasizes scene understanding and categorical reasoning, assessing how well models interpret multi-attribute traffic contexts and safety-critical events.

Dense Captioning Task. For this task, models generate detailed natural-language descriptions of traffic accident videos, capturing key entities, motion, and interactions. The generated captions are compared with ground truth annotations using SPICE [1], METEOR [3], COMET [35], and ROUGE [23] metrics to measure event coverage and linguistic quality. The classic BLEU [32] metric is excluded due to its insensitivity to semantic and contextual variations.

Fine-tuning Task. To assess domain adaptability, we fine-tune Qwen2.5-VL-7B using the VAU-R1 training and evaluation pipeline. The model is reinforcement fine-tuned on the RoadSafe365 QA subset with supervised question–answer pairs, without any additional instruction tuning or multi-stage optimization. We then evaluate the fine-tuned model on the held-out RoadSafe365 test set, enabling a direct comparison between pre-trained and fine-tuned models.

Training and Evaluation. We divide the RoadSafe365 dataset into distinct training and test sets with no overlap, allocating two-thirds for training and one-third for testing. From the training pool, 6,000 accident clips are sampled for fine-tuning, each paired with six attribute-based VQA questions, resulting in 36,000 QA pairs with fixed answer options for consistency. The Qwen2.5-VL-7B model is fine-tuned using the GRPO-based VAU-R1 framework for 3 epochs with a global batch size of 32 (batch size of 1 per GPU with gradient accumulation of 8 across 4 GPUs) and a learning rate of 2×10^{-6}. Gradient checkpointing and FlashAttention-2 are enabled to optimize memory efficiency. Due to computational constraints, we only conduct a pilot study using only a subset of the training data to val-

Table 1. Comparison on the VQA task of RoadSafe365. Columns report accuracy (%) for each attribute (Time, Weather, Area, Road Type, Object, Accident/Violation). Best and second-best are marked in **red** and **blue**.

Models	Year	Attribute Accuracy (%)						**Avg.**
		Time	Weather	Area	Road	Object	Acc./Vio.	
LLaVA-OneVision-7B [20]	2024	92.70	69.89	74.94	39.61	95.69	39.51	68.39
InternVL3-2B [8]	2025	88.44	**75.63**	73.43	66.31	23.54	37.36	60.12
InternVL3-8B [8]	2025	89.77	**75.31**	75.08	**74.07**	99.17	46.31	**76.62**
InternVL3.5-4B [44]	2025	88.16	72.19	65.44	68.15	98.16	**47.91**	73.34
InternVL3.5-8B [44]	2025	78.89	72.51	69.99	72.79	96.24	35.61	70.01
Qwen2.5-VL-3B [2]	2025	87.47	65.63	68.15	51.68	0.50	37.40	51.14
Qwen2.5-VL-7B [2]	2025	**94.77**	66.73	68.98	65.17	35.70	38.04	61.57
Qwen3-VL-4B [2]	2025	**94.26**	72.97	77.56	**74.30**	54.43	47.59	70.52
Qwen3-VL-8B [2]	2025	83.85	74.62	76.09	73.47	69.11	39.88	69.50
LLaVA-NeXT-Video-7B [21]	2024	63.42	53.60	70.90	44.75	**99.86**	25.01	59.25
MiniCPM-V 4.5 [53]	2024	72.74	44.84	**79.76**	30.38	**99.72**	32.63	60.01
Gemini 2.5-flash [9]	2025	83.43	75.26	**78.89**	70.54	88.53	**51.45**	**74.68**

Table 2. Comparison on the dense caption task of RoadSafe365. The best and second performance are marked in **red** and **blue**.

Models	COMET↑	METEOR↑	SPICE↑	ROUGE-1↑			ROUGE-2↑			ROUGE-L↑		
				P	R	F	P	R	F	P	R	F
LLaVA-OneVision-7B [20]	0.641	0.236	0.167	0.386	0.403	0.388	0.103	0.108	0.104	0.228	0.238	0.229
InternVL3-2B [8]	0.707	0.303	0.220	0.450	0.504	0.472	0.145	0.163	0.152	0.256	0.287	0.269
InternVL3-8B [8]	**0.737**	**0.317**	**0.260**	**0.510**	**0.509**	**0.506**	**0.186**	0.184	**0.184**	**0.299**	**0.298**	**0.297**
InternVL3.5-4B [44]	0.736	**0.320**	0.248	0.491	**0.520**	**0.503**	0.176	**0.186**	0.180	0.289	**0.306**	0.295
InternVL3.5-8B [44]	**0.741**	**0.320**	**0.261**	**0.546**	0.503	0.521	**0.201**	**0.185**	**0.192**	**0.318**	0.293	**0.304**
Qwen2.5-VL-3B [2]	0.717	0.266	0.218	0.465	0.436	0.445	0.132	0.123	0.126	0.261	0.244	0.250
Qwen2.5-VL-7B [2]	0.719	0.266	0.221	0.479	0.424	0.447	0.149	0.131	0.139	0.280	0.247	0.261
Qwen3-VL-4B [2]	0.728	0.287	0.225	0.453	0.473	0.461	0.136	0.142	0.138	0.248	0.259	0.252
Qwen3-VL-8B [2]	0.734	0.306	0.222	0.455	0.505	0.476	0.138	0.152	0.144	0.256	0.283	0.267
LLaVA-NeXT-Video-7B [21]	0.667	0.240	0.176	0.408	0.408	0.404	0.122	0.121	0.120	0.255	0.257	0.253
MiniCPM-V 4.5 [53]	0.655	0.274	0.205	0.411	0.443	0.425	0.115	0.123	0.118	0.244	0.263	0.252

idate the effectiveness of fine-tuning on RoadSafe365. Future work is highly encouraged to extend it by utilizing the full training split of RoadSafe365 for large-scale training.

During inference, each video is represented by eight key frames, and all models are evaluated on a held-out test set of 700 clips. We intentionally select a relatively small test set to balance inference time and evaluation quality, making the setup accessible to researchers with limited resources.

4.3. Main Results on Tasks

VQA Task. Table 1 clearly demonstrates that *InternVL3-8B* achieves the highest overall performance (76.62%), followed by *Gemini 2.5-flash* (74.68%) and *InternVL3.5-4B* (73.34%). These models demonstrate strong performance on perception-focused attributes, particularly *Time*, *Weather*, and *Area*, indicating robust visual grounding capabilities. In contrast, most models achieve considerably lower scores on *Accident* and *Violation*, which require causal and safety-related reasoning. Smaller open-source models, such as *Qwen2.5-VL-3B* and *InternVL3-2B*, struggle with these abstract reasoning categories, highlighting their limitations in causal understanding. Overall, perception-oriented attributes (Time, Weather, Area, Road, and Object) typically reach 70–95% accuracy, whereas reasoning-oriented attributes (Acc./Vio.) often fall below 50%. This disparity suggests that while current VLMs can effectively recognize scene conditions, they still face challenges in inferring how and why incidents occur. Additional detailed evaluations of traffic events within the two-level hierarchical structure are provided in the appendix.

Dense Captioning Task. Table 2 presents the dense captioning results on RoadSafe365. *InternVL3.5-8B* achieves the best overall performance across COMET, METEOR, SPICE, and ROUGE, demonstrating strong semantic alignment and well-structured descriptions. *InternVL3-8B* ranks second on most metrics, indicating that larger model scales

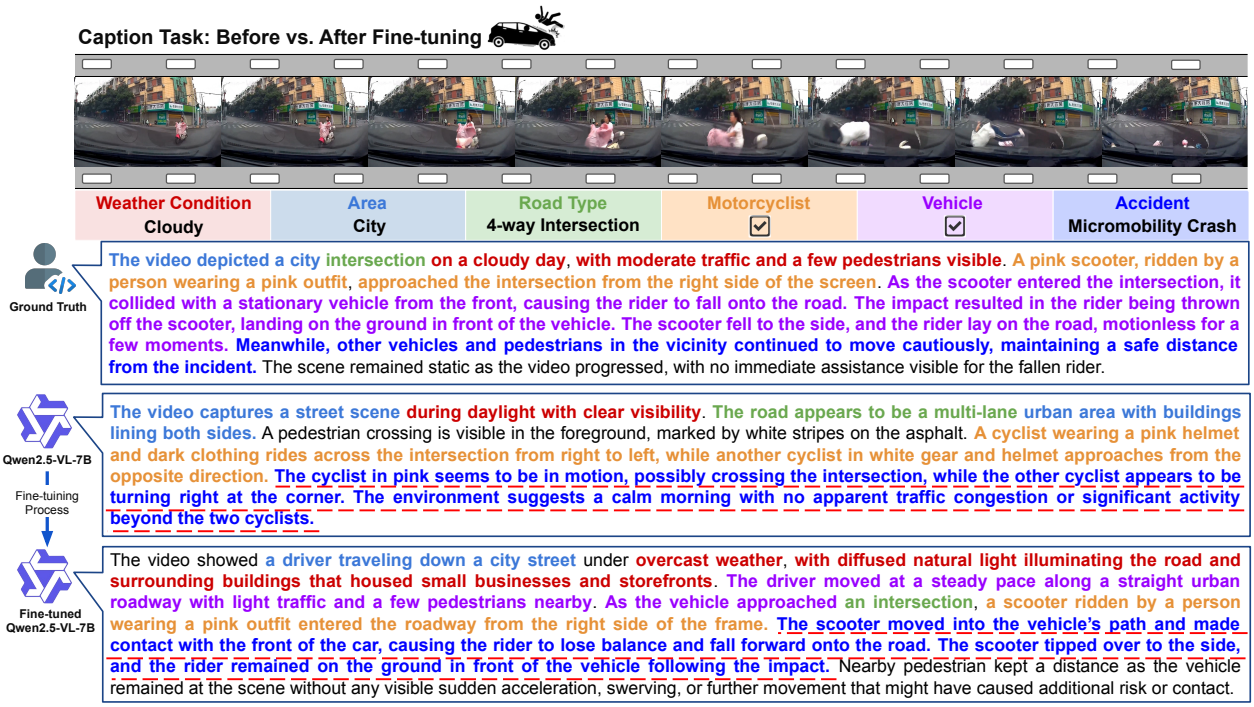

Figure 4. Comparison of accident captions generated by Qwen2.5-VL-7B before and after fine-tuning on RoadSafe365.

further enhance caption quality. Within the Qwen family, *Qwen3-VL-8B* performs the best, generating fluent and detailed captions with the highest COMET and METEOR scores. In contrast, smaller open-source models exhibit clear limitations in capturing accident details and complex interactions. Overall, most models effectively describe scene context and visible actions but struggle with accident dynamics, causal reasoning, and multi-agent interactions.

Fine-Tuning Analysis. We perform reinforcement fine-tuning on Qwen2.5-VL-7B using the VAU-R1 framework with VQA training samples from RoadSafe365, and evaluate the resulting model on RoadSafe365. As shown in Fig. 5, More training iterations consistently improve the VQA performance of Qwen2.5-VL-7B across all categories. Notably, the Accident/Violation subcategory achieves a substantial gain—from 38.04% to 67.29%—an absolute improvement of 29.25%, demonstrating a significant reduction in the gap for safety-critical event understanding. Detailed numbers can be found in the appendix.

Qualitative Results and Discussion. Fig. 4 presents a qualitative comparison of captions generated by Qwen2.5-VL-7B before and after fine-tuning, illustrating how fine-tuning enhances the model's ability to capture causal relationships in traffic accident scenes. Before fine-tuning, the model produced broad scene-level descriptions, emphasizing general context such as weather, road environment, and visible agents. Although fluent and visually grounded, the captions failed to capture the key accident event, overlooked the vehicle–scooter interaction, and did not describe the cause or outcome of the collision. After fine-tuning, the model generated more accurate and coherent accident narratives, it correctly identified the trigger event, namely the scooter rider entering the vehicle's lane and making frontal contact with the car, and clearly explained the cause–effect chain, including the rider's loss of balance, fall to the ground, and

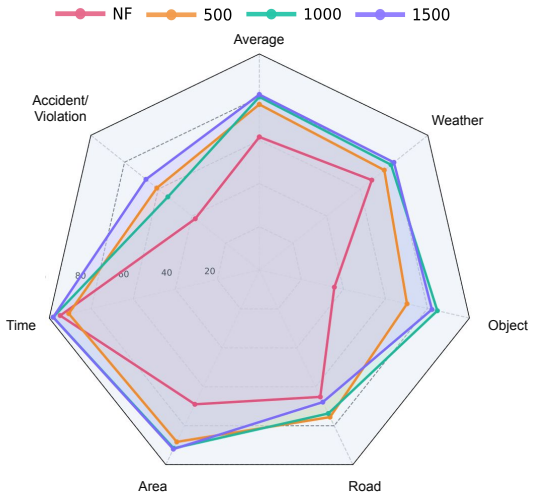

Figure 5. VQA performance of Qwen2.5-VL-7B fine-tuned on RoadSafe365 across different training iterations. NF denotes the original model without reinforcement fine-tuning.

Table 3. Average VQA accuracy (%) of Qwen2.5-VL-7B with different fine-tuning steps across four datasets in the VRU-Accident benchmark [19], namely **DADA**: DADA_2000 [10], **CAP**: CAP_DATA [11], **DoTA**: DoTA [52], **MANU**: MANUAL_DATA.

Model	Steps	DADA	CAP	DoTA	MANU
Qwen2.5-VL-7B	NF	59.27	62.72	57.50	59.79
	900	62.56 ↑3.29	66.09 ↑3.37	59.67 ↑2.17	64.10 ↑4.31
	1200	61.43 ↑2.16	67.02 ↑4.30	61.00 ↑3.50	63.29 ↑3.50
	1500	61.14 ↑1.87	66.61 ↑3.89	60.50 ↑3.00	62.99 ↑3.20

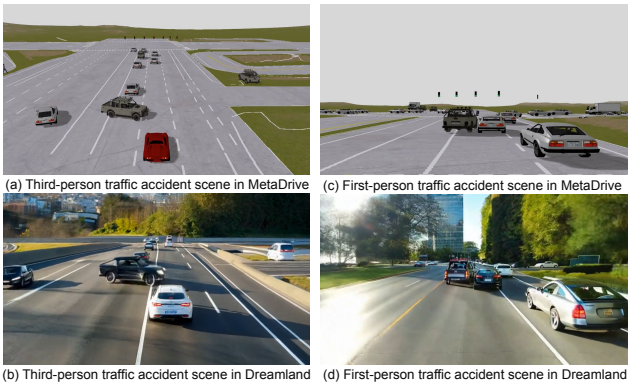

Figure 6. Sample visualizations of synthetic accident scenes from RoadSafe365-Synthetic, created from MetaDrive [22] and Dreamland [27], in both first- and third-person views.

the vehicle stopping immediately afterward. These results show that RoadSafe365 significantly improves the model's causal reasoning, event ordering, and safety-centric understanding.

4.4. Cross-domain Generalization

This section tests RoadSafe365 's cross-domain generalization under two settings: (1) real-world→real-world and (2) real-world→synthetic.

Real-world → Real-world. We reuse the Qwen2.5-VL-7B model fine-tuned via VAU-R1 training and directly evaluate it on the real-world VRU-Accident benchmark [19]. As shown in Table 3, the fine-tuned models at 900, 1200, and 1500 steps consistently outperform the original Qwen2.5-VL-7B model, demonstrating that RoadSafe365 enhances the model's understanding of real accidents—particularly in reasoning and agent interaction. These results further confirm that knowledge learned from RoadSafe365 could effectively transfer to a new real-world accident benchmark.

Real-world → Synthetic. We further construct a realistic synthetic accident dataset via Adv-BMT method [25], into our RoadSafe365-Synthetic, containing two simulation-based sources: (1) MetaDrive [22] clips and (2) Dreamland-rendered [27] scenes, each offering first- and third-person views (Fig. 6). Specifically, we generate 123 third-person and 133 first-person accident clips in MetaDrive, covering diverse road layouts, agent interactions, and accident types. Each scene is then rendered in Dreamland to produce high-fidelity videos with varied lighting conditions and visual styles. All clips are annotated using the RoadSafe365 pipeline, ensuring consistent annotations. Sample annotations could be found in Fig. 7.

Table 4. Average VQA accuracy (%) on synthetic first-person (FV) and third-person (TV) videos, rendered in MetaDrive and Dreamland under different training steps. **NF**: The original Qwen2.5-VL-7B model without reinforcement fine-tuning.

Dataset & View	NF	900	1200	1500
MetaDrive–TV	63.96	71.68 ↑7.72	73.98 ↑10.02	75.61 ↑11.65
MetaDrive–FV	64.41	76.19 ↑11.78	78.57 ↑14.16	78.32 ↑13.91
Dreamland–TV	72.95	77.44 ↑4.49	78.64 ↑5.69	79.16 ↑6.21
Dreamland–FV	67.47	74.89 ↑7.42	75.87 ↑8.40	76.19 ↑8.72

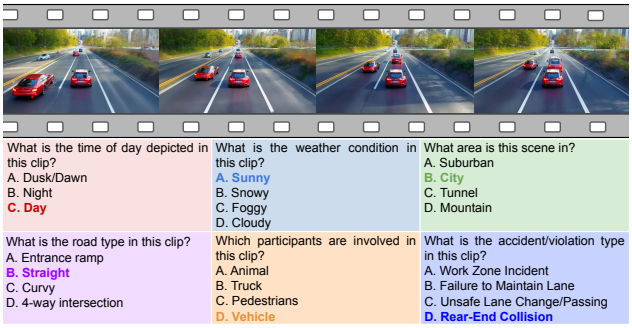

Figure 7. Visualization of annotations from Dreamland [27] in both first-person and third-person views.

We again use the fine-tuned Qwen2.5-VL-7B model and evaluate it on the synthetic dataset. As shown in Table 4, the model consistently surpasses the baseline across all four settings and model snapshots, achieving higher average VQA accuracy throughout.

5. Conclusion

We have introduced RoadSafe365, a large-scale benchmark for real-world traffic safety understanding. It offers fine-grained annotations, a hierarchical taxonomy, and diverse QA tasks that jointly assess the perception, reasoning, and decision-making abilities of VLMs. Extensive experiments in both real-world and synthetic environments show that domain-specific training on RoadSafe365 substantially improves descriptive, reasoning, and generalization performance. RoadSafe365 bridges the gap between visual perception and reasoning in safety-critical environments, providing a unified testbed aligned with official safety standards. We envision this work contributing to the development of explainable, reliable, and human-aligned VLMs for intelligent transportation and autonomous driving.

References

[1] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. Spice: Semantic propositional image caption evaluation. In *European Conference on Computer Vision*, pages 382–398. Springer, 2016. 5

[2] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 5, 6

[3] Satanjeev Banerjee and Alon Lavie. Meteor: an automatic metric for mt evaluation with high levels of correlation with human judgments. *ACL-WMT*, pages 65–72, 2004. 5

[4] Wentao Bao, Qi Yu, and Yu Kong. Drive: Deep reinforced accident anticipation with visual explanation. In *IEEE/CVF International Conference on Computer Vision*, pages 7619–7628, 2021. 1

[5] Digbalay Bose, Rajat Hebbar, Tiantian Feng, Krishna Somandepalli, Anfeng Xu, and Shrikanth Narayanan. Mm-au: Towards multimodal understanding of advertisement videos. In *The 31st ACM International Conference on Multimedia*, pages 86–95, 2023. 3

[6] ByteDance Ltd. Capcut, 2025. Accessed February 10, 2025. 4

[7] Xu Cao, Tong Zhou, Yunsheng Ma, Wenqian Ye, Can Cui, Kun Tang, Zhipeng Cao, Kaizhao Liang, Ziran Wang, James M Rehg, et al. Maplm: A real-world large-scale vision-language benchmark for map and traffic scene understanding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21819–21830, 2024. 2

[8] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24185–24198, 2024. 1, 3, 5, 6

[9] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025. 5, 6

[10] Jianwu Fang, Dingxin Yan, Jiahuan Qiao, Jianru Xue, He Wang, and Sen Li. Dada-2000: Can driving accident be predicted by driver attention*f* analyzed by a benchmark. In *IEEE Intelligent Transportation Systems Conference*, pages 4303–4309. IEEE, 2019. 3, 8, 4

[11] Jianwu Fang, Lei-Lei Li, Kuan Yang, Zhedong Zheng, Jianru Xue, and Tat-Seng Chua. Cognitive accident prediction in driving scenes: A multimodality benchmark. *arXiv preprint arXiv:2212.09381*, 2022. 8, 3, 4

[12] Jianwu Fang, Lei-lei Li, Junfei Zhou, Junbin Xiao, Hongkai Yu, Chen Lv, Jianru Xue, and Tat-Seng Chua. Abductive ego-view accident video understanding for safe driving perception. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22030–22040, 2024. 1

[13] Daniel J Fremont, Tommaso Dreossi, Shromona Ghosh, Xiangyu Yue, Alberto L Sangiovanni-Vincentelli, and Sanjit A Seshia. Scenic: a language for scenario specification and scene generation. In *The 40th ACM SIGPLAN Conference on Programming Language Design and Implementation*, pages 63–78, 2019. 1

[14] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024. 1, 5

[15] Jitesh Jain, Jianwei Yang, and Humphrey Shi. Vcoder: Versatile vision encoders for multimodal large language models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27992–28002, 2024. 1

[16] Muhammad Monjurul Karim, Zhaozheng Yin, and Ruwen Qin. An attention-guided multistream feature fusion network for early localization of risky traffic agents in driving videos. *IEEE Transactions on Intelligent Vehicles*, 9(1):1792–1803, 2023. 3

[17] Aryan Keskar, Srinivasa Perisetla, and Ross Greer. Evaluating multimodal vision-language model prompting strategies for visual question answering in road scene understanding. In *Winter Conference on Applications of Computer Vision*, pages 1027–1036, 2025. 1

[18] Boshra Khalili and Andrew W Smyth. Autodrive-qa-automated generation of multiple-choice questions for autonomous driving datasets using large vision-language models. *arXiv preprint arXiv:2503.15778*, 2025. 3

[19] Younggun Kim, Ahmed S Abdelrahman, and Mohamed Abdel-Aty. Vru-accident: A vision-language benchmark for video question answering and dense captioning for accident scene understanding. In *IEEE/CVF International Conference on Computer Vision*, pages 761–771, 2025. 1, 3, 4, 8

[20] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer. *CoRR*, 2024. 3, 6

[21] Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models. *arXiv preprint arXiv:2407.07895*, 2024. 5, 6

[22] Quanyi Li, Zhenghao Peng, Lan Feng, Qihang Zhang, Zhenghai Xue, and Bolei Zhou. Metadrive: Composing diverse driving scenarios for generalizable reinforcement learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 8

[23] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004. 5

[24] Shaoyu Liu, Jianing Li, Guanghui Zhao, Yunjian Zhang, Xin Meng, Fei Richard Yu, Xiangyang Ji, and Ming Li. Event-gpt: Event stream understanding with multimodal large language models. In *Computer Vision and Pattern Recognition Conference*, pages 29139–29149, 2025. 1

[25] Yuxin Liu, Zhenghao Peng, Xuanhao Cui, and Bolei Zhou. Adv-bmt: Bidirectional motion transformer for safety-critical traffic scenario generation. *arXiv preprint arXiv:2506.09485*, 2025. 8

[26] Haohan Luo and Feng Wang. A simulation-based framework for urban traffic accident detection. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 1–5. IEEE, 2023. 3

[27] Sicheng Mo, Ziyang Leng, Leon Liu, Weizhen Wang, Honglin He, and Bolei Zhou. Dreamland: Controllable world creation with simulator and generative models. *arXiv preprint arXiv:2506.08006*, 2025. 8, 10

[28] Daniel Moura, Shizhan Zhu, and Orly Zvitia. Nexar dashcam collision prediction dataset and challenge. In *Computer Vision and Pattern Recognition Conference*, pages 2583–2591, 2025. 1

[29] National Highway Traffic Safety Administration. Early estimate of motor vehicle traffic fatalities in 2024. Technical Report DOT HS 813 710, National Highway Traffic Safety Administration, 2025. 1

[30] National Safety Council. Committee on Uniform Traffic Accident Statistics and American Standards Association. *Manual on classification of motor vehicle traffic accidents*. National Safety Council, 1962. 1, 4

[31] Chenbin Pan, Burhaneddin Yaman, Tommaso Nesti, Abhirup Mallik, Alessandro G Allievi, Senem Velipasalar, and Liu Ren. Vlp: Vision language planning for autonomous driving. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14760–14769, 2024. 1, 3

[32] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *The 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, 2002. 5

[33] Chirag Parikh, Deepti Rawat, Tathagata Ghosh, Ravi Kiran Sarvadevabhatla, et al. Roadsocial: A diverse videoqa dataset and benchmark for road event understanding from social video narratives. In *Computer Vision and Pattern Recognition Conference*, pages 19002–19011, 2025. 3

[34] Sung-Yeon Park, Can Cui, Yunsheng Ma, Ahmadreza Moradipari, Rohit Gupta, Kyungtae Han, and Ziran Wang. Nuplanqa: A large-scale dataset and benchmark for multi-view driving scene understanding in multi-modal large language models. *CoRR*, 2025. 2

[35] Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. Comet: A neural framework for mt evaluation. In *Conference on Empirical Methods in Natural Language Processing*, pages 2685–2702, 2020. 5

[36] Hao Shao, Yuxuan Hu, Letian Wang, Guanglu Song, Steven L Waslander, Yu Liu, and Hongsheng Li. Lmdrive: Closed-loop end-to-end driving with large language models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15120–15130, 2024. 3

[37] Ankit Kumar Shaw, Chandan Kumar Sah, Xiaoli Lian, Arsalan Shahid Baig, Tuopu Wen, Kun Jiang, Mengmeng Yang, Diange Yang, and Li Zhang. Saferoute: Enhancing traffic scene understanding via a unified deep learning and multimodal llm. In *IEEE/CVF International Conference on Computer Vision*, pages 4547–4556, 2025. 1

[38] Zihao Sheng, Zilin Huang, Yansong Qu, Yue Leng, and Sikai Chen. Talk2traffic: Interactive and editable traffic scenario generation for autonomous driving with multimodal large language model. In *Computer Vision and Pattern Recognition Conference*, pages 3788–3797, 2025. 1

[39] Chonghao Sima, Katrin Renz, Kashyap Chitta, Li Chen, Hanxue Zhang, Chengen Xie, Jens Beißwenger, Ping Luo, Andreas Geiger, and Hongyang Li. Drivelm: Driving with graph visual question answering. In *European Conference on Computer Vision*, 2024. 3

[40] Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024. 1

[41] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *CoRR*, 2024. 1, 3

[42] Shihao Wang, Zhiding Yu, Xiaohui Jiang, Shiyi Lan, Min Shi, Nadine Chang, Jan Kautz, Ying Li, and Jose M Alvarez. Omnidrive: A holistic vision-language dataset for autonomous driving with counterfactual reasoning. In *Computer Vision and Pattern Recognition Conference*, pages 22442–22452, 2025. 1

[43] Weizhen Wang, Chenda Duan, Zhenghao Peng, Yuxin Liu, and Bolei Zhou. Embodied scene understanding for vision language models via metavqa. In *Computer Vision and Pattern Recognition Conference*, pages 22453–22464, 2025. 1, 2

[44] Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025. 6

[45] Hsiu-Fu Wu, Ya-Ting Yang, Yung-Ter Chen, and I-Fan Chou. Trafficinternvl: Understanding traffic scenarios with vision-language models. In *IEEE/CVF International Conference on Computer Vision*, pages 5229–5236, 2025. 1

[46] Shaoyuan Xie, Lingdong Kong, Yuhao Dong, Chonghao Sima, Wenwei Zhang, Qi Alfred Chen, Ziwei Liu, and Liang Pan. Are vlms ready for autonomous driving? an empirical study from the reliability, data, and metric perspectives. *CoRR*, 2025. 1, 3

[47] Shuo Xing, Hongyuan Hua, Xiangbo Gao, Shenzhe Zhu, Renjie Li, Kexin Tian, Xiaopeng Li, Heng Huang, Tianbao Yang, Zhangyang Wang, et al. Autotrust: Benchmarking trustworthiness in large vision language models for autonomous driving. *CoRR*, 2024. 3

[48] Li Xu, He Huang, and Jun Liu. Sutd-trafficqa: A question answering benchmark and an efficient network for video reasoning over traffic events. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9878–9888, 2021. 3

[49] Zhenhua Xu, Yan Bai, Yujia Zhang, Zhuoling Li, Fei Xia, Kwan-Yee K Wong, Jianqiang Wang, and Hengshuang Zhao. Drivegpt4-v2: Harnessing large language model capabilities for enhanced closed-loop autonomous driving. In *Computer Vision and Pattern Recognition Conference*, pages 17261–17270, 2025. 3

[50] Xuemeng Yang, Licheng Wen, Tiantian Wei, Yukai Ma, Jianbiao Mei, Xin Li, Wenjie Lei, Daocheng Fu, Pinlong Cai, Min Dou, et al. Drivearena: A closed-loop generative simulation platform for autonomous driving. In *IEEE/CVF International Conference on Computer Vision*, pages 26933–26943, 2025. 3

[51] Yu Yao, Mingze Xu, Yuchen Wang, David J Crandall, and Ella M Atkins. Unsupervised traffic accident detection in first-person videos. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 273–280. IEEE, 2019. 3

[52] Yu Yao, Xizi Wang, Mingze Xu, Zelin Pu, Yuchen Wang, Ella Atkins, and David J Crandall. Dota: Unsupervised detection of traffic anomaly in driving videos. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(1): 444–459, 2022. 8, 3, 4

[53] Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, et al. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*, 2024. 6

[54] Tackgeun You and Bohyung Han. Traffic accident benchmark for causality recognition. In *European Conference on Computer Vision*, pages 540–556. Springer, 2020. 1, 3

[55] Xingcheng Zhou, Konstantinos Larintzakis, Hao Guo, Walter Zimmer, Mingyu Liu, Hu Cao, Jiajie Zhang, Venkatnarayanan Lakshminarasimhan, Leah Strand, and Alois C Knoll. Tumtraffic-videoqa: A benchmark for unified spatio-temporal video understanding in traffic scenes. *arXiv preprint arXiv:2502.02449*, 2025. 3

[56] Yixuan Zhou, Long Bai, Sijia Cai, Bing Deng, Xing Xu, and Heng Tao Shen. Tau-106k: A new dataset for comprehensive understanding of traffic accident. In *The Thirteenth International Conference on Learning Representations*, 2025. 3, 4

[57] Liyun Zhu, Qixiang Chen, Xi Shen, and Xiaodong Cun. Vau-r1: Advancing video anomaly understanding via reinforcement fine-tuning. *arXiv preprint arXiv:2505.23504*, 2025. 5

[58] Walter Zimmer, Ross Greer, Daniel Lehmberg, Marc Pavel, Holger Caesar, Xingcheng Zhou, Ahmed Ghita, Mohan Trivedi, Rui Song, Hu Cao, et al. Towards vision zero: The accid3nd dataset. *arXiv preprint arXiv:2503.12095*, 2025. 3