RoadSocial: A Diverse VideoQA Dataset and Benchmark for Road Event Understanding from Social Video Narratives

Chirag Parikh*, Deepti Rawat*, Rakshitha R. T., Tathagata Ghosh, and Ravi Kiran Sarvadevabhatla
CVIT & iHub-Data, IIIT Hyderabad, India

https://roadsocial.github.io

Abstract

We introduce RoadSocial, a large-scale, diverse VideoQA dataset tailored for generic road event understanding from social media narratives. Unlike existing datasets limited by regional bias, viewpoint bias and expert-driven annotations, RoadSocial captures the global complexity of road events with varied geographies, camera viewpoints (CCTV, handheld, drones) and rich social discourse. Our scalable semi-automatic annotation framework leverages Text LLMs and Video LLMs to generate comprehensive question-answer pairs across 12 challenging QA tasks, pushing the boundaries of road event understanding. RoadSocial is derived from social media videos spanning 14M frames and 414K social comments, resulting in a dataset with 13.2K videos, 674 tags and 260K high-quality QA pairs. We evaluate 18 Video LLMs (open-source and proprietary, driving-specific and general-purpose) on the RoadSocial-QA benchmark. We also demonstrate RoadSocial's utility in improving road event understanding capabilities of general-purpose Video LLMs.

1. Introduction

A road event typically refers to any incident, activity, or condition occurring on or around the roadway that affects traffic flow, safety, or road usage. The ability to recognize and interpret road events is essential for safe and reliable intelligent vehicles and transportation systems. In this regard, large-scale video datasets of road events are used to develop assistive models [2, 3, 7, 18, 21, 25]. Many recent datasets contain videos with accompanying question-answer text pairs and other text metadata [13, 19, 20, 23]. Such datasets have become a de facto choice for training Video Large Language Models (Video LLMs) [6, 11, 23, 33].

However, current video-based road event understanding approaches are limited by region-specific datasets, neglecting the diversity of global road scenarios. Most datasets focus on dashcam views for autonomous driving, overlooking other camera types such as CCTV, handheld, and drone-based. They also lack annotations on generic events (*e.g.* defensive driving, near-misses). Due to the reliance on regionally-biased expert annotators, the broader and richer contextual insights from real-world social discourse on road events are absent. Furthermore, existing evaluation frameworks fail to test the Video LLMs' ability to distinguish informative road event details from misleading information, essential for developing reliable, hallucination-resistant road event understanding systems.

To address these limitations and to enable foundational video language models for *generic* road event understanding, we introduce **RoadSocial**, a large-scale and diverse Video Question Answer (VideoQA) dataset. RoadSocial is obtained by processing social media videos and the narratives accompanying these videos. The inherent diversity of social media in terms of geographical locations, camera viewpoints, road event types and social commentary addresses shortcomings of video datasets mentioned previously. Specifically, we make the following contributions:

- RoadSocial: a large-scale, diverse VideoQA resource for road events, derived from social media videos spanning **14M** frames and **414K** social comments from **100** countries, resulting in a dataset with **13.2K** videos, **674** unique tags, and **260K** high-quality QA pairs.
- A semi-automatic annotation framework using Text LLM and Video LLM that processes social media video narratives and generates comprehensive QA pairs across 12 distinct challenging tasks.
- A robust evaluation framework incorporating *non-road event* videos and irrelevant questions to assess the robustness of Video LLMs to hallucinations.
- A demonstration of RoadSocial's utility in improving road event understanding capabilities of general-purpose VideoLLM and for evaluating Planning/AV-related tasks.
- Critical insights into 18 Video LLMs' performance on road event understanding, obtained from their evaluation

*Equal contribution.

on our RoadSocial-QA benchmark.

2. RoadSocial Dataset

RoadSocial is a dataset created from social media videos in unconstrained, real-world environments. These videos are accompanied by rich social commentary that reflects facts and varied cultural perspectives on road events worldwide.

2.1. Data Collection

We crowdsourced diverse road event data from X (formerly Twitter), leveraging its global community for real-world insights. Unlike other platforms, X is characterized by an active social discourse on road events that includes the general public, road event enthusiasts, and road enforcement authorities. Our strategy focused on popular road event related social media handles worldwide, using multilingual keywords to scrape tweet data from 2012 onwards, filtering for videos with substantial commentary. The resulting dataset captures varied road events—traffic violations, accidents, safe driving, and infrastructure awareness—across different environments and locations.

2.2. Annotation Strategy: QAs and Tags

Our annotation strategy leverages both visual content and social media narratives to generate high-quality, culturally diverse QA pairs and video tags (Figs. 1 and 2).

Identifying Representative Road Event Samples: We start by identifying representative samples of diverse road events observed across different geographical regions and design effective template questions for QA generation. We extract vector embeddings of multilingual tweet text and hashtags using OpenAI's GPT-3 text embeddings [17]. We perform hierarchical k-means clustering over the embeddings to produce clusters of distinct road events (*e.g.* UK cyclist near-misses, illegal truck overtaking in China, car hydroplaning in USA). We selected top five samples closest to each cluster center as representatives, ensuring our QA generation is grounded in well-represented road events.

Hybrid Approach for QA Generation: We then use a hybrid approach to generate QA pairs that blend video semantics with social media context. We split videos into 3-second segments and use VideoLLMs to generate detailed captions, which are then processed by TextLLMs to create cohesive visual summary. Tweet conversations are then cleaned and integrated with visual summary to generate contextually rich QA pairs using template questions. Generated QA pairs undergo refinement to remove non-visual information, and ensure human-like responses. Finally, all QA pairs are verified by expert annotators. We also generate incompatible QA pairs for non-road event videos to test model robustness to hallucinations, and adversarial QA pairs to evaluate model's ability to reject misleading assumptions.

Video-level Tag Generation: To categorize videos by key aspects of road events, we generate diverse video-level tags (*e.g.* traffic violation, wheelie, unsafe overtaking) using the verified answers discussed earlier. A Text LLM [1] scans these answers to generate top-k tags most relevant to each QA pair of unique type defined by the template questions. This structured tagging approach ensures that the generated QA pairs, tags are robust and reflect the diverse scenarios present in the dataset. Distribution of the resulting tags for each QA pair type are shown by word clouds in Fig. 1.

2.3. Dataset Statistics

Our final dataset stats are shown in Fig. 2. The dataset exhibits significant diversity across several dimensions, including geographical distribution (Fig. 2), QA types (Fig. 3), and video tags with their distribution shown in word clouds (Fig. 1). The global coverage of our dataset attributes, depicts the diverse cultural perspectives involved in the QA pair generation process. It includes **414K** multilingual tweet captions and replies corresponding to **204K** unique user handles (from across **100** countries) sharing facts and opinions about the road or traffic events.

2.4. QA Tasks Taxonomy

We developed a question-answer (QA) taxonomy for structured evaluation of Video Large Language Models (Video LLMs). The taxonomy consists of 12 distinct tasks organized into four reasoning categories: Complex, Factual, Imaginative, and Hallucination (Fig. 3). Our taxonomy extends beyond conventional road datasets by incorporating previously underrepresented tasks, such as Viewpoint QA (analysis of camera perspectives) and Where QA (geographic location identification). As an additional novelty, our approach uniquely incorporates Adversarial QA and Incompatible QA to test models' robustness to hallucinations and rejecting misleading assumptions or false details in questions.

3. Experiments and Analysis

To evaluate zero-shot reasoning capabilities of Video LLMs, we split our dataset into **12K** training and **1.2K** test videos, resulting in **234K** and **26K** QA pairs respectively. The video splits maintain geographical diversity across the dataset, with the test set serving as our primary evaluation benchmark. We evaluated 18 Video LLMs (both open-source and proprietary, driving-specific and general-purpose) on our benchmark using video frames with task specific questions. The results of their zero-shot performance on all QA tasks is presented in Tab. 1.

Following established practices in recent literature [10, 12, 23, 29], we report GPT-3.5 score [15] for all QA tasks

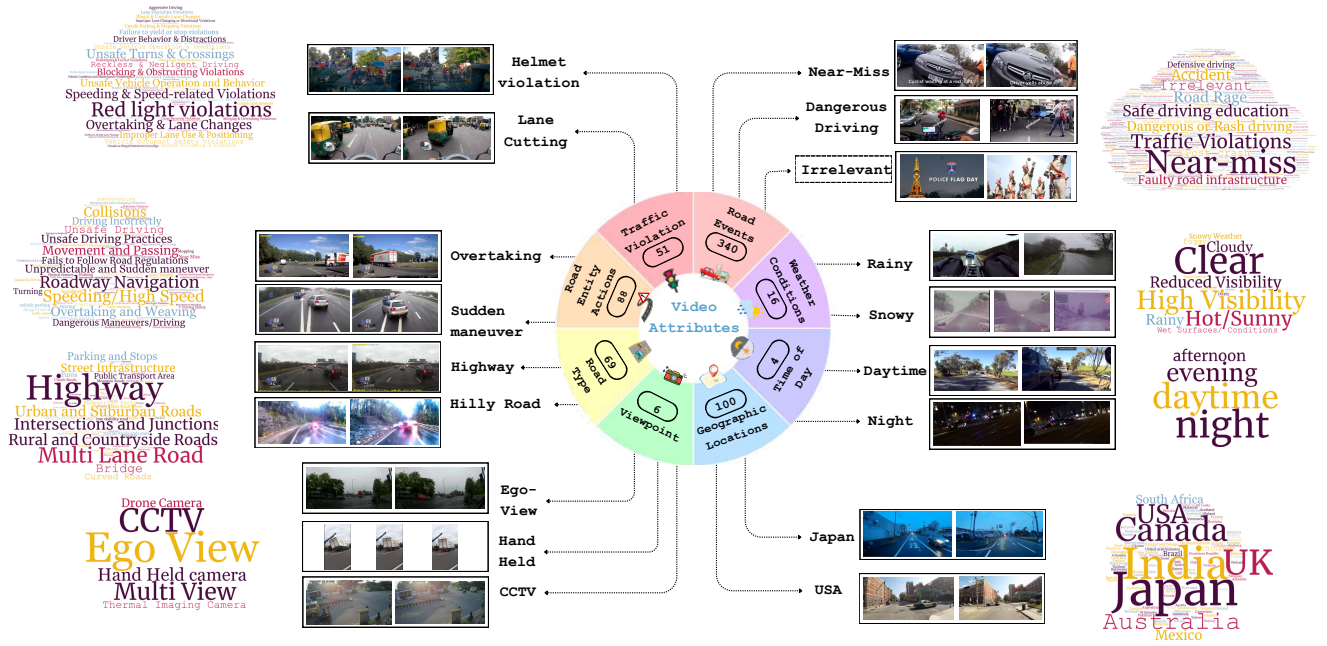

Figure 1. **Diverse Video Attributes in the RoadSocial Dataset:** The total count of unique tags for each attribute is shown in circled boxes, alongside word clouds highlighting these values. For each attribute, we display examples with 2-3 keyframes from videos. The figure captures the diversity of road events, environmental conditions, geographical locations, viewpoints, interactions between road entities, and traffic violations. The varied scenarios under each attribute showcase the rich complexity of our dataset.

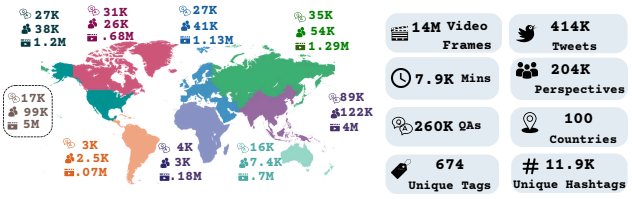

Figure 2. **The diversity of RoadSocial dataset:** The number of QA pairs, social commentary (tweets), and video frames spread across different regions is shown. Overall statistics of the raw tweet data, generated QA pairs, and tags in our dataset is also shown. Total incompatible QA pairs and related numbers for non-road event videos are specified inside a light brown box at left.

Figure 3. **QA Task Taxonomy:** The QA pairs in RoadSocial are broadly grouped into 4 categories (highlighted in blue) which are further subdivided into 12 tasks (shown in green). Total QA pair count for each category is shown in blue squared box. Some of these tasks are further subdivided into granular sub-tasks (highlighted in orange) to facilitate coarse to fine-grained understanding of road events along different aspects.

(except Temporal Grounding) to assess the similarity between model-generated and ground-truth open-ended responses. For Temporal Grounding QAs, mean Average Precision (mAP) score is reported which compares time interval ranges in model-generated and ground-truth responses.

Key findings include: (1) All general-purpose models surpass driving-specific Video LLM across all QA tasks, revealing the regional bias in driving-focused models causing performance gaps in general road event understanding. (2) Models struggle more with generic questions than specific ones due to the need for context during inference. (3) Strong performance in factual reasoning (Where, Viewpoint) but weakness in complex reasoning (Description, Why). (4) Temporal grounding of road events is particularly challenging (<1% mAP for most models). (5) Promising capabilities in imaginative reasoning (>70% GPT-scores in Advisory, Introspection tasks). (6) Variable robustness to hallucination with few models effectively rejecting misleading information (Incompatible, Adversarial).

We demonstrate in Tab. 1 (last row) that fine-tuning a

Model	Params	Factual			Complex				Imaginative			Hallucination		Overall	Overall	Overall	Overall
		WR	KE	VP	DS	WY	CQ	TG	AD	IN	CF	AV	IC	(ALL)	(RT)	(Generic)	(Specific)
Dolphin [11]	9B	61.3	34.5	67.8	35.8	25.2	37.2	0.01	49.8	39.1	45.5	71.8	21.3	40.8	42.5	29.8	46.5
GPT-4o [16]	-	77.0	**66.6**	84.3	**70.2**	**70.8**	**72.1**	7.8	**77.7**	**76.4**	**77.0**	**90.0**	**67.6**	**69.8**	**70.0**	**69.5**	**74.4**
Gemini-1.5-Pro [26]	-	**77.7**	56.7	**85.4**	61.9	61.4	60.1	**18.6**	72.1	70.2	75.7	72.3	48.7	63.4	64.7	60.1	68.3
InternVL2 [4]	76B	72.4	51.3	81.4	57.1	59.0	62.1	1.07	70.5	67.0	69.2	58.6	27.6	56.4	59.1	55.5	65.1
Qwen2-VL [28]	72B	76.8	56.6	85.1	60.2	64.0	67.6	0.01	71.9	72.4	71.6	37.0	40.2	58.6	60.3	58.3	68.8
LLaVA-Video [32]	72B	75.8	52.4	76.8	52.4	55.0	52.2	**9.94**	68.3	63.7	64.9	83.5	24.7	56.7	59.6	51.1	63.3
LLaVA-OV [8]	72B	75.1	54.1	78.7	53.0	53.3	54.1	3.99	67.8	61.9	63.1	45.1	19.9	52.5	55.5	51.8	63.0
VITA [5]	8x7B	66.6	52.1	71.6	48.1	55.6	56.3	2.27	66.7	66.0	62.4	56.3	22.0	52.2	54.9	49.8	60.4
Tarsier [27]	34B	73.7	58.1	78.2	58.2	59.0	58.8	0.32	71.6	71.1	67.4	83.2	**82.3**	**63.5**	61.8	58.4	66.1
ARIA [9]	25.3B	75.4	53.1	**86.2**	58.4	56.9	**70.2**	8.96	**75.1**	74.7	74.0	**86.4**	29.2	62.4	65.4	56.7	68.5
InternVL2 [4]	8B	67.7	51.7	78.0	55.7	59.3	60.9	0.77	66.7	66.8	70.0	68.1	26.1	56.0	58.7	53.7	64.0
Mini-CPM-V 2.6 [30]	8B	77.7	57.6	80.6	55.0	50.5	57.5	0.4	61.6	52.3	59.3	73.5	30.0	54.7	56.9	51.0	62.0
IXC-2.5 [31]	7B	**78.5**	**58.7**	85.4	**61.7**	**65.3**	68.5	0.69	73.9	**75.6**	**75.7**	85.8	29.2	63.3	**66.4**	**60.7**	**70.3**
Tarsier [27]	7B	69.9	54.7	72.3	52.0	53.4	55.2	0.11	69.5	69.3	63.5	79.1	67.3	58.9	58.1	54.0	61.7
LongVU [22]	7B	73.0	53.0	76.3	51.1	50.2	55.0	0.84	59.7	55.8	58.2	48.9	32.7	51.2	52.9	47.7	59.7
Qwen2-VL [28]	7B	75.5	52.8	76.1	52.7	57.7	56.4	0.59	69.2	71.6	65.9	37.5	39.6	54.6	56.0	52.6	63.9
LLaVA-Video [32]	7B	74.6	50.1	76.7	52.1	50.1	50.3	1.43	60.4	53.8	58.7	61.8	23.5	51.1	53.6	47.6	59.7
LLaVA-OV [8]	7B	73.4	51.2	77.2	50.7	51.7	51.2	0.97	62.8	55.4	58.6	45.4	21.1	50.0	52.6	48.4	59.8
LLaVA-OV ft.	7B	80.9	64.1	85.7	64.1	68.7	65.1	4.49	74.2	70.9	71.7	95.4	87.6	69.4	67.8	65.1	69.7

Table 1. **Video LLMs benchmarked on RoadSocial-QA.** Standard prompting with task-specific instructions were employed for zero-shot evaluation of Video LLMs on 12 QA tasks. Video LLMs are grouped as open-source (driving-specific and general-purpose), and closed-source models. Further, we fine-tune a Video LLM - LLaVA-OV-7B and report its performance at the end of the table. Abbreviations used for QA tasks include Factual (F), Complex (C), Imaginative (I), Hallucination (H), Where (WR), Key Entities (KE), Viewpoint (VP), Description (DS), Why (WY), Consequence (CQ), Temporal Grounding (TG), Advisory (AD), Introspection (IN), Counterfactual (CF), Adversarial (AV), Incompatible (IC), and Road-event related Tasks (RT). RT includes all tasks except IC which corresponds to non-road event videos. GPT-3.5 score is reported for all tasks except Temporal Grounding (TG) for which average mAP@.3:.7 (%) is reported. Overall average scores are reported for ALL QA tasks (F, C, I, and H), Road-event related Tasks (RT), Generic QAs, and Specific QAs under each task. All reported scores (scale 0 to 100) are colored based on their value from low to high. VideoLLMs show per-query latencies of 1-25s (7B-76B) on H100 GPUs.

general-purpose VideoLLM [8] on our dataset yields a substantial 19.4% improvement across all QA tasks, highlighting RoadSocial's effectiveness in enhancing general road event understanding. Furthermore, Tab. 2 shows that the fine-tuned VideoLLM achieves significant gains on popular autonomous driving benchmarks—specifically, the PlanningQA task in DriveLM [24] and the Action/Scenery QA task in Lingo-QA [14]. These results further demonstrate RoadSocial's utility in improving the understanding of video-based planning and AV-related tasks.

Model	DriveLM Planning [24]	Lingo-QA Eval [14]
LLaVA-OV [8]	31.7	37.0
LLaVA-OV ft.	40.1 (+8.3%)	41.6 (+4.6%)

Table 2. Performance comparison of LLaVA-OV [8] with and without fine-tuning on RoadSocial dataset.

4. Conclusion

RoadSocial redefines the landscape for general-purpose road event understanding. By capturing diverse camera viewpoints, geographical contexts, and socially-informed QAs, RoadSocial delivers a comprehensive dataset that captures the complexity of real-world road scenarios across varied cultural and environmental contexts. Leveraging social media content, it addresses the limitations of traditional datasets by incorporating unique perspectives and nuanced social discourse. Our scalable semi-automatic annotation framework, powered by Text and Video LLMs, can easily ingest and process social media posts generated continuously over time, enabling even larger dataset size with sustained quality. Our robust evaluation framework tests model resilience to irrelevant inputs, hallucinations, cross-viewpoint comprehension, and geographical awareness. Our evaluation across 18 Video LLMs provides critical performance insights across a spectrum of road event QA tasks. We believe RoadSocial will be instrumental in driving progress towards safer and more inclusive intelligent transportation systems.

References

[1] Anthropic. Claude 3.5 model. https://www.

anthropic.com/news/claude-3-5-sonnet, 2024. Accessed: 2024-11-08. 2

[2] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11621–11631, 2020. 1

[3] Rohan Chandra, Xijun Wang, Mridul Mahajan, Rahul Kala, Rishitha Palugulla, Chandrababu Naidu, Alok Jain, and Dinesh Manocha. Meteor: A dense, heterogeneous, and unstructured traffic dataset with rare behaviors. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 9169–9175. IEEE, 2023. 1

[4] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Zhong Muyan, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. Intern vl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 24185–24198, 2023. 4

[5] Chaoyou Fu, Haojia Lin, Zuwei Long, Yunhang Shen, Meng Zhao, Yifan Zhang, Xiong Wang, Di Yin, Long Ma, Xiawu Zheng, et al. Vita: Towards open-source interactive omni multimodal llm. *arXiv preprint arXiv:2408.05211*, 2024. 4

[6] Akshay Gopalkrishnan, Ross Greer, and Mohan M. Trivedi. Multi-frame, lightweight & efficient vision-language models for question answering in autonomous driving. *ArXiv*, abs/2403.19838, 2024. 1

[7] Jinkyu Kim, Anna Rohrbach, Trevor Darrell, John Canny, and Zeynep Akata. Textual explanations for self-driving vehicles. In *Proceedings of the European conference on computer vision (ECCV)*, pages 563–578, 2018. 1

[8] 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. *arXiv preprint arXiv:2408.03326*, 2024. 4

[9] Dongxu Li, Yudong Liu, Haoning Wu, Yue Wang, Zhiqi Shen, Bowen Qu, Xinyao Niu, Guoyin Wang, Bei Chen, and Junnan Li. Aria: An open multimodal native mixture-of-experts model. *arXiv preprint arXiv:2410.05993*, 2024. 4

[10] Stephanie Lin, Jacob Hilton, and Owain Evans. Truthfulqa: Measuring how models mimic human falsehoods. *arXiv preprint arXiv:2109.07958*, 2021. 2

[11] Yingzi Ma, Yulong Cao, Jiachen Sun, Marco Pavone, and Chaowei Xiao. Dolphins: Multimodal language model for driving. *arXiv preprint arXiv:2312.00438*, 2023. 1, 4

[12] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. *arXiv preprint arXiv:2306.05424*, 2023. 2

[13] Ana-Maria Marcu, Long Chen, Jan Hünermann, Alice Karnsund, Benoit Hanotte, Prajwal Chidananda, Saurabh Nair, Vijay Badrinarayanan, Alex Kendall, Jamie Shotton, et al. Lingoqa: Visual question answering for autonomous driving. *arXiv preprint arXiv:2312.14115*. 1

[14] Ana-Maria Marcu, Long Chen, Jan Hünermann, Alice Karnsund, Benoit Hanotte, Prajwal Chidananda, Saurabh Nair, Vijay Badrinarayanan, et al. Lingoqa: Visual question answering for autonomous driving. In *ECCV*, 2024. 4

[15] OpenAI. Openai: Introducing chatgpt. https://openai.com/blog/chatgpt/, 2022. 2

[16] OpenAI. Hello gpt-4o. https://openai.com/index/hello-gpt-4o/, 2024. 4

[17] OpenAI. Vector embeddings. https://platform.openai.com/docs/guides/embeddings, 2024. Accessed: 2024-11-08. 2

[18] Chirag Parikh, Rohit Saluja, C. V. Jawahar, and Ravi Kiran Sarvadevabhatla. Idd-x: A multi-view dataset for ego-relative important object localization and explanation in dense and unstructured traffic. *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 14815–14821, 2024. 1

[19] Ehsan Qasemi, Jonathan M Francis, and Alessandro Oltramari. Traffic-domain video question answering with automatic captioning. *ArXiv*, abs/2307.09636, 2023. 1

[20] Tianwen Qian, Jingjing Chen, Linhai Zhuo, Yang Jiao, and Yu-Gang Jiang. Nuscenes-qa: A multi-modal visual question answering benchmark for autonomous driving scenario. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 4542–4550, 2024. 1

[21] Vasili Ramanishka, Yi-Ting Chen, Teruhisa Misu, and Kate Saenko. Toward driving scene understanding: A dataset for learning driver behavior and causal reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7699–7707, 2018. 1

[22] Xiaoqian Shen, Yunyang Xiong, Changsheng Zhao, Lemeng Wu, Jun Chen, Chenchen Zhu, Zechun Liu, Fanyi Xiao, Balakrishnan Varadarajan, Florian Bordes, et al. Longvu: Spatiotemporal adaptive compression for long video-language understanding. *arXiv preprint arXiv:2410.17434*, 2024. 4

[23] Chonghao Sima, Katrin Renz, Kashyap Chitta, Li Chen, Hanxue Zhang, Chengen Xie, Ping Luo, Andreas Geiger, and Hongyang Li. Drivelm: Driving with graph visual question answering. *arXiv preprint arXiv:2312.14150*, 2023. 1, 2

[24] 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 *ECCV*, 2024. 4

[25] Gurkirt Singh, Stephen Akrigg, Manuele Di Maio, Valentina Fontana, Reza Javanmard Alitappeh, Salman Khan, Suman Saha, Kossar Jeddisaravi, Farzad Yousefi, Jacob Culley, et al. Road: The road event awareness dataset for autonomous driving. *IEEE transactions on pattern analysis and machine intelligence*, 45(1):1036–1054, 2022. 1

[26] Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023. 4

[27] Jiawei Wang, Liping Yuan, and Yuchen Zhang. Tarsier: Recipes for training and evaluating large video description models. *ArXiv*, abs/2407.00634, 2024. 4

[28] 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. *arXiv preprint arXiv:2409.12191*, 2024. 4

[29] Zhenhua Xu, Yujia Zhang, Enze Xie, Zhen Zhao, Yong Guo, Kwan-Yee. K. Wong, Zhenguo Li, and Hengshuang Zhao. Drivegpt4: Interpretable end-to-end autonomous driving via large language model. *IEEE Robotics and Automation Letters*, 9:8186–8193, 2023. 2

[30] 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. 4

[31] Pan Zhang, Xiao wen Dong, Yuhang Zang, Yuhang Cao, Rui Qian, Lin Chen, Qipeng Guo, Haodong Duan, Bin Wang, Linke Ouyang, Songyang Zhang, Wenwei Zhang, Yining Li, Yang Gao, Peng Sun, Xinyue Zhang, Wei Li, Jingwen Li, Wenhai Wang, Hang Yan, Conghui He, Xingcheng Zhang, Kai Chen, Jifeng Dai, Yu Qiao, Dahua Lin, and Jiaqi Wang. Internlm-xcomposer-2.5: A versatile large vision language model supporting long-contextual input and output. *ArXiv*, abs/2407.03320, 2024. 4

[32] Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Video instruction tuning with synthetic data, 2024. 4

[33] Yunsong Zhou, Linyan Huang, Qingwen Bu, Jia Zeng, Tianyu Li, Hang Qiu, Hongzi Zhu, Minyi Guo, Yu Qiao, and Hongyang Li. Embodied understanding of driving scenarios. *arXiv preprint arXiv:2403.04593*, 2024. 1