V2X-based Logical Scenario Understanding with Vision-Language Models

Cheng-Liang Chi Zi-Hui Li Yu-Hsiang Chen Yi-Ting Chen
National Yang Ming Chiao Tung University

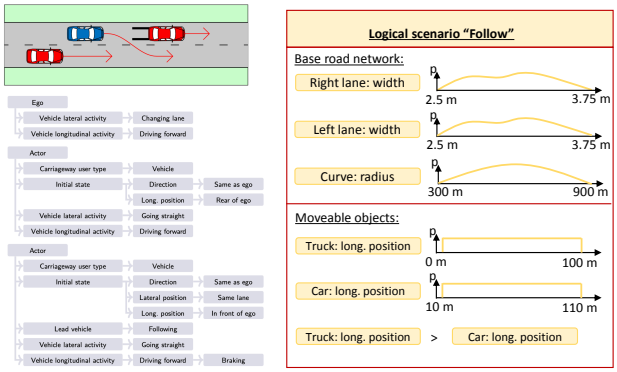
Abstract

Ensuring autonomous vehicle (AV) safety requires testing across diverse traffic scenarios, yet real-world data collection is costly and rarely captures safety-critical events. Simulation enables scalable evaluation, but defining and annotating realistic scenarios remains a challenge. Logical scenarios are parameterized representations of traffic situations that specify agent behavior and environmental context, enabling structured and repeatable testing across a wide range of conditions. We construct a multi-view dataset by combining scenario templates with varied parameters and rendering them from ego, drone, and infrastructure viewpoints. To support scalable high-level annotation, we evaluate several Large Vision-Language Models (LVLMs) on the task of video-level scenario categorization. While most models struggle with structured classification, performance improves for select models when using prompting strategies and aerial views. Our findings reveal both the limitations and emerging potential of LVLMs in supporting semantic scenario understanding for autonomous driving.

1. Introduction

Validating autonomous vehicle (AV) safety is challenging because critical events are rare in real-world data and difficult to observe systematically. Large-scale testing across diverse locations, conditions, and rare scenarios is expensive, time-consuming, and often impractical. Simulation offers a safer, more scalable alternative, especially for evaluating V2X-based cooperative systems that combine infrastructure, onboard, and aerial perception. Scenario-based methods enable systematic evaluation by partitioning the operational design domain (ODD) into parameterized traffic situations defined by constraints on environment, agent behavior, and road geometry.

Scenario-based simulation enables targeted testing of specific configurations and edge cases under controlled conditions, but defining parameterized logical scenarios that reflect the real world presents significant challenges. Determining parameter ranges that reflect real-world diversity and frequency is difficult without extensive real-world data.



(a) Lead vehicle braking and ego lane change. Picture credit: [13].

(b) Parametrized following scenario. Picture credit: [23].

Figure 1. Examples of logical scenarios. (a) illustrates semantic composition using behavioral tags, and (b) shows numerical parameter ranges for road geometry and agent configuration.

Examples of logical scenarios and their parameterizations are illustrated in Fig. 1, highlighting both semantic composition and numerical parameter specification.

To define realistic parameter ranges for simulation, it is necessary to capture the complex dynamics of real-world traffic. Ego-centric, single-view perception is limited by occlusion and narrow field of view, making cooperative V2X perception essential. By fusing data from heterogeneous sensors, V2X systems create high-fidelity representations of traffic scenes. Extracting parameter distributions from continuous V2X video requires accurate spatiotemporal calibration across sensors [34] and robust retrieval methods to ground entities and events in raw multimodal data [12].

Scene annotation for scenario-based validation relies on two complementary label types. Low-level annotations, including agent positions, velocities, and trajectories, support parameter extraction and dominate existing datasets [7, 14, 27], which focus on detection, tracking, and motion prediction. High-level semantic labels instead describe the overall traffic situation, such as a vehicle turning at an intersection or merging into a lane. These enable logical scenario classification, crucial for structured testing and systematic ODD coverage [13, 28]. However, manual labeling is time-consuming and hard to scale. This work aims to automate high-level annotation with LVLMs to support

scalable scenario-based validation in autonomous driving.

To address scalable semantic annotation, we propose leveraging LVLMs for high-level scenario understanding. Unlike traditional classifiers that require retraining for new categories, LVLMs trained on large, diverse datasets enable zero-shot generalization to unseen scenarios without fine-tuning. This reduces the manual effort and cost of expanding scenario taxonomies. Recent LVLMs, including proprietary models like GPT-4o [1] and open-source variants such as Qwen-VL [5], show strong vision-language reasoning performance. We explore their use for video-level classification of traffic scenes into logical scenarios, aiming to enable scalable semantic annotation for scenario-based validation in autonomous driving.

2. Related Work

2.1. Visual Understanding of Traffic Scenes

Visual scene understanding has evolved from image classification and object detection to more structured and dense prediction tasks. While models such as ResNet [15], DETR [10], and Mask R-CNN [16] have advanced object-level perception, they offer limited support for holistic traffic scene interpretation. Semantic, instance, and panoptic segmentation further improve granularity, yet rely heavily on dense annotations and task-specific architectures, constraining generalization across diverse environments.

To better capture relationships between entities, recent approaches incorporate relational modeling. Scene graph generation methods like RelTR [11] introduce relational representations between objects. In traffic domains, recent work has focused on modeling interactions and dynamics [32]. For instance, Action Slot [18] leverages slot attention to capture motion-aware, object-centric features, while DAGCN [19] performs action recognition using 3D pose-based relational reasoning. Other efforts reconstruct road topology from egocentric images [9], translating onboard views into structured bird's-eye-view graphs that incorporate semantics, object locations, and connectivity [8, 20].

Despite progress in perception and relational modeling, existing methods struggle with generalization in diverse traffic conditions. Scenario categorization offers a high-level understanding of traffic scenes by abstracting agent behaviors and context into semantic labels. We investigate the use of vision-language foundation models to automate this process in cooperative V2X settings, enabling scalable annotation without task-specific supervision.

2.2. Vision-Language Models

While task-specific perception models have made significant progress, they often struggle to generalize to long-tail and complex real-world traffic scenarios. Recent studies have shown that standard object detection models exhibit severe performance drops when evaluated on in-the-wild traffic datasets, highlighting a persistent generalization gap [2]. These challenges motivate the use of more flexible, data-efficient models that can adapt to diverse visual contexts with minimal supervision.

Vision-Language Models (VLMs) offer this flexibility through large-scale multimodal pretraining, enabling open-vocabulary recognition, zero-shot generalization, and high-level reasoning. Recent models extend these capabilities to video. Video-LLaMA [30] introduces a Q-Former for aligning temporal and audio-visual features with language. Video-LLaVA [21] unifies image and video inputs into a shared representation. InternLM-XComposer [31] treats video as high-resolution composite images for dense temporal reasoning. Qwen2.5-VL [4] incorporates dynamic frame-rate processing and absolute time encoding for long-range event localization. Cosmos-Reason1 [25] adds physics-aware reasoning via supervised fine-tuning and reinforcement learning.

Collectively, these VLMs introduce new capabilities that go beyond detection or segmentation, offering interpretable and adaptable perception suitable for open-world driving environments. Their ability to infer semantic roles, model agent interactions, and generalize across unseen contexts positions them as a foundational component for robust and scalable autonomous driving systems.

2.3. Benchmarks for Vision-Language Models in Traffic Scenes

Despite growing interest in applying VLMs to driving, existing benchmarks fail to capture the temporal and behavioral complexity of real-world traffic. Most rely on static, single-frame inputs, overlooking the sequential context essential for dynamic scenes. Video datasets such as DRAMA [22] provide annotated driving clips with spatio-temporal events but lack scenario categorization and support for high-level reasoning. TUMTraffic-VideoQA [33] extends evaluation to video question answering but remains limited to short-span queries without structured behavioral coverage. Other efforts [24] emphasize temporal consistency but omit agent behavior, while nuScenes-QA [26] frames understanding as QA without addressing scene classification or interaction modeling. These gaps highlight the need for temporally grounded, behavior-rich benchmarks to evaluate VLMs in complex traffic environments.

3. Methodology

3.1. Dataset Construction

We collected a V2X dataset by manually defining 20 basic logical scenarios, each with predefined parameter ranges and involving only a single agent. Complex scenarios are constructed by combining these basic scenarios with var-

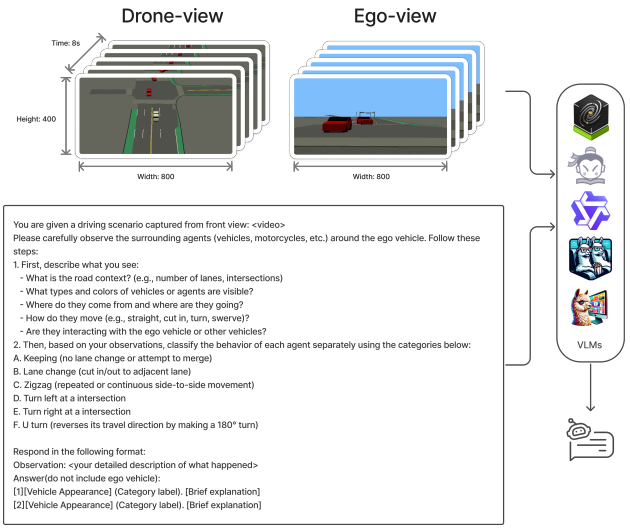


Figure 2. Evaluation pipeline for different VLMs. The order from top to bottom are Cosmos-Reason1 [25], InternLM-XComposer-2.5 [31], Qwen2.5-VL [4], VideoLLaMA 2 [30], Video-LLaVA [21]. The input is a video of a traffic scene, either ego-front view or drone-view, and the output is the predicted agent behavior. The model is prompted to reason about the scene and classify the agent's behavior into one of six categories.

ied parameters, resulting in 9K OpenSCENARIO-format files. We use esmini (an OpenSCENARIO player) [17] to replay these scenarios while simultaneously synthesizing novel multi-view video data.

Each scenario is rendered from 10 distinct viewpoints: six surrounding the ego-vehicle, three drone perspectives at different angle, and one fixed infrastructure camera view.

Because each scenario is built from predefined logical templates, we have ground-truth annotations describing agent behavior. Initially, we defined 20 behavior categories, but due to semantic overlap and difficulty in clearly distinguishing some of them using natural language, we consolidated them into six representative categories: keeping, turn left, turn right, lane change, zigzag, and U-turn.

3.2. Models and Prompt Engineering

We adopt Qwen2.5-VL [4] and Cosmos-Reason1 [25] as our primary VLMs for traffic scene understanding. Qwen2.5-VL is a recently released model that achieves state-of-the-art performance across multiple video understanding benchmarks. Unlike LLaVA-based models such as Video-LLaVA [21], Qwen2.5-VL does not suffer from severe token limitations, making it compatible with in-context learning [6] and long-format reasoning.

We also experimented with several other popular VLMs, including VideoLLaMA 2 [30], Video-LLaVA [21], and InternLM-XComposer-2.5 [31]. However, these models exhibited major limitations in instruction-following and structured response generation. Despite carefully designed prompts and demonstrations, they failed to produce classification results in the expected format. For example, Video-LLaVA frequently generated repetitive or irrelevant responses, and could not handle even simple Visual Question Answering (VQA) [3] questions such as "Is there a car in front?". The responses were often inconsistent or overlay generic, e.g., "The car is driving on the road and is in the middle of the road," with little variation or semantic depth for traffic scenes, whereas non-traffic images were described in more detail. Video-LLaMA-2 exhibited similar behavior, which we attribute to overfitting on their instruction-tuning datasets that lack sufficient coverage of structured traffic scene data and classification tasks.

In contrast, Qwen2.5-VL demonstrates more reliable instruction-following and produces coherent, scene-relevant descriptions in traffic scenarios. We therefore choose it as our base model. Furthermore, Cosmos-Reason1, built upon Qwen2.5-VL, is additionally trained on datasets related to the physical world, such as robotics and traffic-centric reasoning. We hypothesize that this additional grounding enhances its capability to model complex agent interactions in real-world driving scenes.

To leverage these models effectively, we design task-specific prompts that incorporate two key strategies: Chain-of-Thought (CoT) [29] and In-Context Learning (ICL) [6]. For CoT prompting, we guide the model through a step-by-step reasoning process, first describing the agents in the scene and their behaviors before requesting a final classification. For ICL, we provide one or more exemplars in the prompt that demonstrate the expected input-output format, enabling the model to generalize to new traffic scenes without further tuning.

4. Experimental Results

We evaluate multiple vision-language foundation models on diverse driving scenarios using different input viewpoints. The experiments focus on two core comparisons: (1) Comparing ego-view with drone-view inputs to assess the impact of spatial perspective on scene understanding. (2) Evaluating different model performance with varying in-context learning (ICL) examples, particularly under ego-view conditions.

4.1. Viewpoint Comparison: Ego vs. Drone

We compare ego-centric and drone views using Qwen2.5-VL-7B with vehicle-only (VO) and full-scene (Full) inputs, as shown in Fig. 3. In all configurations, the drone-based view consistently outperforms the ego-front view, particularly under the 2-shot and 4-shot settings. This suggests that the global spatial awareness provided by the bird's-eye perspective enables more effective reasoning about traffic scenarios, even without direct agent-centric cues.

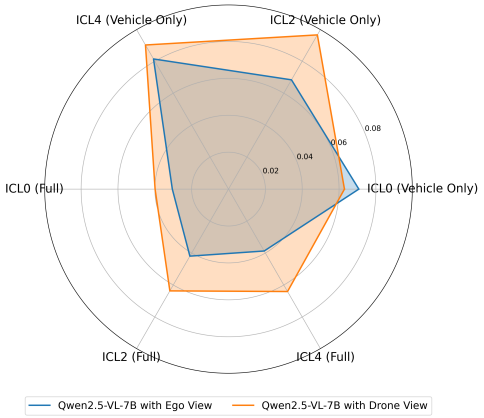


Figure 3. Comparison of ego-centric and drone views using Qwen2.5-VL-7B under VO and Full dataset settings, with varying ICL examples. The metrics are weighted F1 scores across six behavior categories.

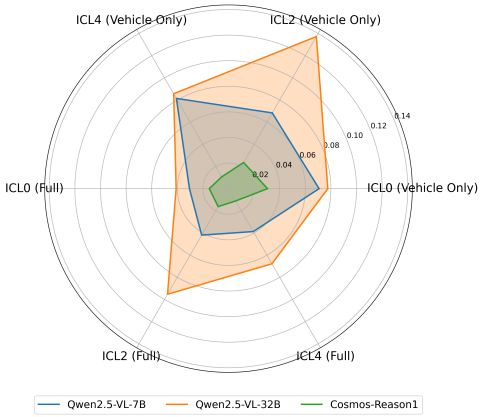


Figure 4. Comparison of different LVLMs on ego-front view inputs, evaluating performance across VO and Full dataset settings with varying ICL examples. The metrics are weighted F1 scores across six behavior categories.

Moreover, we observe that increasing the number of in-context examples generally improves performance in both ego and drone views. In most cases, the weighted F1 scores increase from 0-shot to 2-shot, suggesting that few-shot prompting enables the model to better interpret complex scenes through demonstration-based reasoning. However, the improvement from 2-shot to 4-shot is sometimes marginal, indicating that a small number of relevant examples may already suffice, and that additional prompts beyond a certain point offer diminishing returns.

4.2. Model Comparison under Ego-Front View

We evaluate the performance of different models, Qwen2.5-VL-7B, Qwen2.5-VL-32B, and Cosmos-Reason1-7B, using the ego-front view, as shown in Fig. 4. The analysis focuses on how each model handles the complexities of driving scenarios under both vehicle-only (VO) and full-scene (Full) input settings, across different in-context learning configurations (ICL0, ICL2, ICL4).

Cosmos-Reason1-7B demonstrates consistently poor performance across all configurations. A qualitative analysis reveals that the model frequently fails to follow the intended classification instructions, instead reverting to behaviors aligned with its instruction-tuned training objectives. For instance, even without prompts related to temporal alignment, Cosmos often generates outputs that attempt to identify frame sequences rather than describing scenarios or providing class predictions. This suggests a strong tendency to overfit to its post-training dataset, which heavily emphasizes video alignment and retrieval tasks.

In contrast, both Qwen2.5-VL variants exhibit more reliable alignment with the classification objective. Overall, configurations with in-context learning (ICL2 and ICL4) outperform the 0-shot CoT baseline, reaffirming that few-shot prompting contributes positively to performance. The larger Qwen2.5-VL-32B model consistently outperforms its 7B counterpart, indicating that increased model capacity enhances the ability to leverage contextual examples. However, for Qwen2.5-VL-32B, we observe that performance at ICL4 is slightly lower than at ICL2 in both VO and Full scenarios. This suggests that a small number of well-designed exemplars may already saturate the model's ability to generalize, and that additional shots may introduce unnecessary noise or distract from the core reasoning process.

5. Conclusion

In this paper, we presented benchmarks for evaluating VLMs in complex traffic scenarios, focusing on multi-agent interactions and behavioral classification. Our dataset, constructed from a diverse set of logical scenarios, provides a rich foundation for assessing VLMs' capabilities in understanding dynamic traffic environments. Our findings highlight the potential of VLMs to enhance scenario-based validation pipelines for autonomous driving systems.

6. Future Work

Future work will extend the dataset and evaluation framework to better support V2X cooperative perception. This includes constructing multi-viewpoint scenarios and integrating additional sensing modalities such as LiDAR and radar for cross-agent fusion and benchmarking.

The current evaluation focuses on behavior classification. To assess scene understanding more comprehensively, future metrics should evaluate fine-grained grounding, including the ability to localize and distinguish agents by spatial configuration, appearance, and context. Additionally, fine-tuning models on the dataset and comparing them with existing traffic scene VQA benchmarks will help assess generalization and dataset utility.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 2

[2] Sara A. Al-Emadi, Yin Yang, and Ferda Ofli. Benchmarking object detectors under real-world distribution shifts in satellite imagery. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2025. 2

[3] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Int. Conf. Comput. Vis.*, 2015. 3

[4] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 2, 3

[5] 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. 2

[6] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *Adv. Neural Inform. Process. Syst.*, 2020. 3

[7] Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yuxin Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2020. 1

[8] Yigit Baran Can, Alexander Liniger, Danda Pani Paudel, and Luc Van Gool. Structured bird's-eye-view traffic scene understanding from onboard images. In *Int. Conf. Comput. Vis.*, 2021. 2

[9] Yigit Baran Can, Alexander Liniger, Danda Pani Paudel, and Luc Van Gool. Topology preserving local road network estimation from single onboard camera image. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2022. 2

[10] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *Eur. Conf. Comput. Vis.*, 2020. 2

[11] Yuren Cong, Michael Ying Yang, and Bodo Rosenhahn. Reltr: Relation transformer for scene graph generation. In *IEEE Trans. Pattern Anal. Mach. Intell.*, 2023. 2

[12] Cainan Davidson, Deva Ramanan, and Neehar Peri. Refav: Towards planning-centric scenario mining. *arXiv preprint arXiv:2505.20981*, 2025. 1

[13] Erwin de Gelder, Olaf Op den Camp, and Niels de Boer. Scenario categories for the assessment of automated vehicles. *CETRAN, Singapore, Version*, 1, 2020. 1

[14] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are 1we ready for autonomous driving? the kitti vision benchmark suite. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2012. 1

[15] Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2016. 2

[16] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross B. Girshick. Mask r-cnn. *Int. Conf. Comput. Vis.*, 2017. 2

[17] Emil Knabe and Melih Guldogus. esmini - openscenario player and simulator. https://github.com/esmini/esmini, 2019. 3

[18] Chi-Hsi Kung, Shu-Wei Lu, Yi-Hsuan Tsai, and Yi-Ting Chen. Action-slot: Visual action-centric representations for multi-label atomic activity recognition in traffic scenes. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2024. 2

[19] T. Li, Z. Zhao, C. Sun, R. Yan, and X. Chen. Domain adversarial graph convolutional network for fault diagnosis under variable working conditions. *IEEE Transactions on Instrumentation and Measurement*, 2021. 2

[20] Zhiqi Li, Wenhai Wang, Hongyang Li, Enze Xie, Chonghao Sima, Tong Lu, Qiao Yu, and Jifeng Dai. Bevformer: learning bird's-eye-view representation from lidar-camera via spatiotemporal transformers. In *IEEE Trans. Pattern Anal. Mach. Intell.*, 2024. 2

[21] Bin Lin, Bin Zhu, Yang Ye, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection. In *Empirical Methods in Natural Language Processing*, 2024. 2, 3

[22] Srikanth Malla, Chiho Choi, Isht Dwivedi, Joon Hee Choi, and Jiachen Li. Drama: Joint risk localization and captioning in driving. In *IEEE Winter Conf. Appl. Comput. Vis.*, 2023. 2

[23] Till Menzel, Gerrit Bagschik, and Markus Maurer. Scenarios for development, test and validation of automated vehicles. In *2018 IEEE Intelligent Vehicles Symposium (IV)*, 2018. 1

[24] Athma Narayanan, Isht Dwivedi, and Behzad Dariush. Dynamic traffic scene classification with space-time coherence. In *IEEE Int. Conf. Robot. Autom.*, 2019. 2

[25] NVIDIA, Alisson Azzolini, Hannah Brandon, Prithvijit Chattopadhyay, Huayu Chen, Jinju Chu, Yin Cui, Jenna Diamond, Yifan Ding, Francesco Ferroni, Rama Govindaraju, Jinwei Gu, Siddharth Gururani, Imad El Hanafi, Zekun Hao, Jacob Huffman, Jingyi Jin, Brendan Johnson, Rizwan Khan, George Kurian, Elena Lantz, Nayeon Lee, Zhaoshuo Li, Xuan Li, Tsung-Yi Lin, Yen-Chen Lin, Ming-Yu Liu, Andrew Mathau, Yun Ni, Lindsey Pavao, Wei Ping, David W. Romero, Misha Smelyanskiy, Shuran Song, Lyne Tchapmi, Andrew Z. Wang, Boxin Wang, Haoxiang Wang, Fangyin Wei, Jiashu Xu, Yao Xu, Xiaodong Yang, Zhuolin Yang, Xiaohui Zeng, and Zhe Zhang. Cosmos-reason1: From physical common sense to embodied reasoning, 2025. 2, 3

[26] 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 *AAAI*, 2024. 2

[27] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, Vijay Vasudevan, Wei Han, Jiquan Ngiam, Hang Zhao, Aleksei Timofeev, Scott Ettinger, Maxim Krivokon, Amy Gao, Aditya Joshi, Yu Zhang, Jonathon Shlens, Zhifeng Chen, and Dragomir Anguelov. Scalability in perception for autonomous driving: Waymo open dataset. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2020. 1

[28] Hendrik Weber, Julian Bock, Jens Klimke, Christian Rösener, Johannes Hiller, Robert Krajewski, Adrian Zlocki, and Lutz Eckstein. A framework for definition of logical scenarios for safety assurance of automated driving. *Traffic Injury Prevention*, 2019. 1

[29] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Adv. Neural Inform. Process. Syst.*, 2022. 3

[30] Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. In *Empirical Methods in Natural Language Processing*, 2023. 2, 3

[31] Pan Zhang, Xiaoyi 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 preprint arXiv:2407.03320*, 2024. 2, 3

[32] Jing Zhao, Victor L. Knoop, and Meng Wang. Microscopic traffic modeling inside intersections: Interactions between drivers. *Transportation Science*, 57, 2023. 2

[33] 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. In *International Conference on Machine Learning*, 2025. 2

[34] Zewei Zhou, Hao Xiang, Zhaoliang Zheng, Seth Z. Zhao, Mingyue Lei, Yun Zhang, Tianhui Cai, Xinyi Liu, Johnson Liu, Maheswari Bajji, Jacob Pham, Xin Xia, Zhiyu Huang, Bolei Zhou, and Jiaqi Ma. V2xpnp: Vehicle-to-everything spatio-temporal fusion for multi-agent perception and prediction. In *Int. Conf. Comput. Vis.*, 2025. 1