Title => Abstract => Conclusion => Expriment Result => Introduction => Method

TRAFFIC-MLLM: A SPATIO-TEMPORAL MLLM WITH RETRIEVAL-AUGMENTED GENERATION FOR CAUSAL INFERENCE IN TRAFFIC

🟥 problem

🟨 highlight

🟩 improvement

Waikit Xiu, Qiang Lu, Xiying Li, Chen Hu, Shengbo Sun

Sun Yat-sen University
School of Intelligent Systems Engineering

ABSTRACT

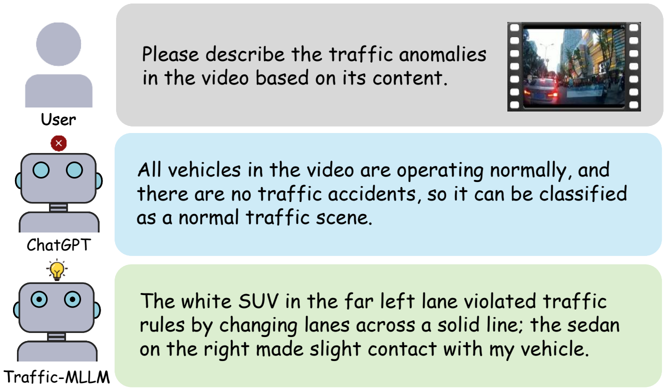
As intelligent transportation systems advance, traffic video understanding plays an increasingly pivotal role in comprehensive scene perception and causal analysis. Yet, existing approaches face notable challenges in accurately modeling spatiotemporal causality and integrating domain-specific knowledge, limiting their effectiveness in complex scenarios. To address these limitations, we propose Traffic-MLLM, a multimodal large language model tailored for fine-grained traffic analysis. Built on the Qwen2.5-VL backbone, our model leverages high-quality traffic-specific multimodal datasets and uses Low-Rank Adaptation (LoRA) for lightweight fine-tuning, significantly enhancing its capacity to model continuous spatiotemporal features in video sequences. Furthermore, we introduce an innovative knowledge prompting module fusing Chain-of-Thought (CoT) reasoning with Retrieval-Augmented Generation (RAG), enabling precise injection of detailed traffic regulations and domain knowledge into the inference process. This design markedly boosts the model's logical reasoning and knowledge adaptation capabilities. Experimental results on TrafficQA and DriveQA benchmarks show Traffic-MLLM achieves state-of-the-art performance, validating its superior ability to process multimodal traffic data. It also exhibits remarkable zero-shot reasoning and cross-scenario generalization capabilities.

***Index Terms*—** Multimodal Large Language Models, Traffic Video Understanding, Spatio-temporal Reasoning

1. INTRODUCTION

Computer vision has made significant advances in transportation, supporting tasks like vehicle detection[1] and infrastructure assessment[2, 3], thereby promoting intelligent traffic development. However, task-specific approaches struggle to meet the complexity and variability of real-world traffic scenarios. As intelligent transportation systems evolve, the industry increasingly demands comprehensive scene understanding, making the development of end-to-end multimodal large models a key focus.

Current research in the transportation domain includes numerous large-scale models[4, 5], primarily focused on big data time-series forecasting[6] and traffic event text analysis driven by large language models[7]. However, for cross-modal tasks involving traffic videos and images, mainstream approaches predominantly rely on visual-text contrastive learning. These methods exhibit significant limitations: their capacity to model spatiotemporal dependencies is inadequate, as they typically match static image features, discretized video frames, and labels, failing to capture continuous spatiotemporal dynamics necessary for understanding multi-object interactions and ensuring temporal coherence. Additionally, these approaches suffer from weak generalization and overfitting, stemming from limited knowledge transfer and data adaptation capabilities; they depend heavily on scene-specific annotations and tend to overfit local features. Although general multimodal large models such as Qwen-VL[8] and Gemini Pro Vision[9] possess some ability for spatiotemporal modeling and cross-scenario transfer, their insufficient adaptation to fine-grained domain knowledge leads to issues akin to "model hallucinations" in traffic applications. This highlights the critical need for developing specialized multimodal large models tailored to the unique requirements of the traffic domain.



Please describe the traffic anomalies in the video based on its content.

All vehicles in the video are operating normally, and there are no traffic accidents, so it can be classified as a normal traffic scene.

The white SUV in the far left lane violated traffic rules by changing lanes across a solid line; the sedan on the right made slight contact with my vehicle.

Fig. 1. Surpassing General-Purpose Understanding: Traffic-MLLM Achieves Fine-Grained Traffic Violation Recognition

To address the core challenges, we propose Traffic-MLLM, a multimodal large model specifically designed for fine-grained traffic image and video understanding in complex traffic scenarios. The model is built upon Qwen2.5VL and leverages high-quality traffic image and video annotation datasets, along with a lightweight LoRA[10] fine-tuning strategy, to effectively enhance perception of static details and spatiotemporal features of continuous frames, thereby tackling the issue of insufficient spatiotemporal causal modeling. Additionally, the model features an innovative traffic knowledge prompt enhancement module that integrates Chain-of-Thought (CoT)[11] reasoning and Retrieval-Augmented Generation (RAG) mechanisms: CoT guides the model to systematically analyze causal relationships in complex traffic scenes, while RAG dynamically retrieves relevant knowledge from traffic regulations and scene standards, injecting fine-grained domain knowledge to mitigate "hallucination" problems caused by domain knowledge gaps. Experimental results demonstrate that Traffic-MLLM per-

LLM - context engineer -> thêm infor sau mi ln suy lun

suy lun cn truy vn

RAG giúp truy vn -> úng ng cnh, ngn context

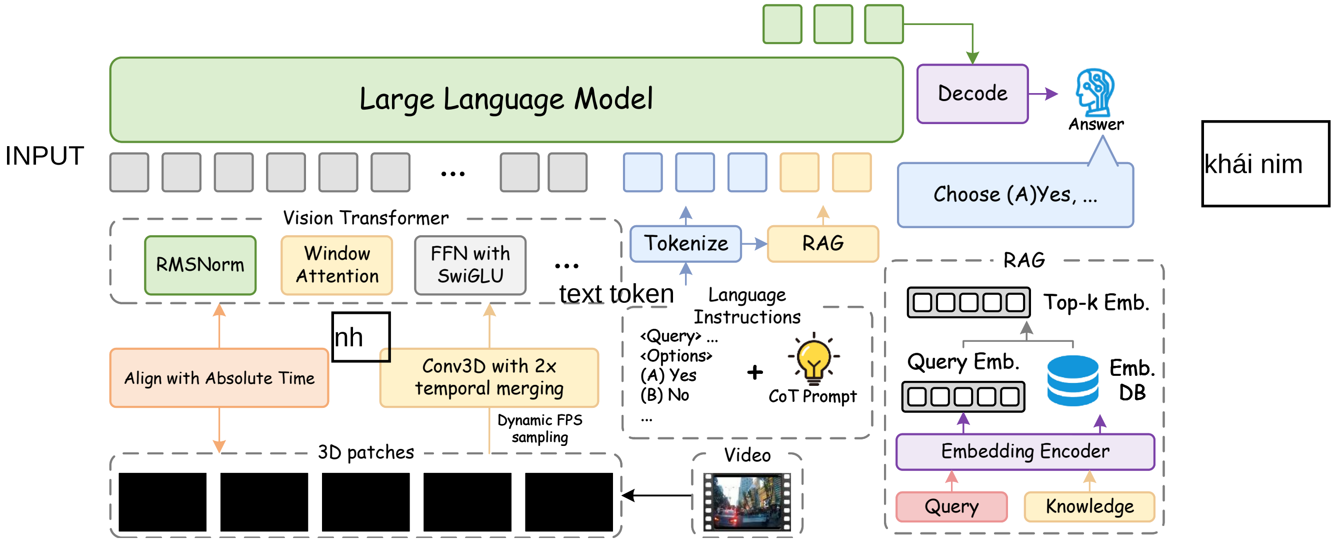


Fig. 2. Overall architecture of the retrieval-augmented multimodal reasoning framework for traffic scenarios.

forms excellently on traffic multimodal understanding benchmarks such as TrafficQA[12] and DriveQA[13], achieving state-of-the-art performance.

Therefore, the main contributions of this work can be summarized as follows: (1) We propose Traffic-MLLM, a dedicated multimodal model for traffic scenarios, which effectively addresses the lack of domain-specific adaptation in general-purpose models and offers a reference for developing vertical-domain models; (2) We design a traffic knowledge prompt enhancement module that significantly improves the model's ability to incorporate professional traffic knowledge with low computational overhead; (3) We validate the model's superior performance on authoritative benchmarks, achieving excellent results on TrafficQA and state-of-the-art (SOTA) zero-shot reasoning performance on DriveQA among models of comparable scale.

The overall structure is as follows, Section 2 details the methodology, Section 3 presents experiments and analysis, and Section 4 concludes.

2. METHODOLOGY

In this section, we introduce Traffic-MLLM, a model for understanding fine-grained vehicle behavior and conducting causal reasoning from traffic videos. Section 2.1 presents the model's framework, Section 2.2 discusses enhancements from Trans-RAG for causal reasoning, and Section 2.3 outlines our dataset construction and model training pipeline.

2.1. Traffic MLLM Architecture

In this section, we introduce the framework of our proposed method. We adopt Qwen2.5-VL-3B as our base model and explore strategies to enhance its capabilities in traffic event understanding and causal reasoning.

Visual Encoder: The visual encoder employs an enhanced Vision Transformer (ViT) architecture. It first performs 3D patching on video frames, followed by spatiotemporal encoding. For spatiotemporal position encoding, 2D Rotary Position Embedding (2D-RoPE) is applied to encode the spatial coordinates of each patch, which is then combined with MRoPE embeddings aligned with absolute time to capture two-dimensional spatial relationships. The internal network uses RMSNorm as the normalization function, formulated as eq(1):

$$\mathrm{RMSNorm}(x) = \frac{x}{\sqrt{\mathrm{mean}(x^2) + \epsilon}} \times \gamma \tag{1}$$

where x is the input vector, $\mathrm{mean}(x^2)$ is the mean of the squares of the elements in x, ϵ is a small constant for numerical stability, and γ is a learnable scaling parameter. SwiGLU is employed as the activation function, with its Feed-Forward Network (FFN) formulation given by eq(2):

$$\mathrm{SwiGLU}(x) = \mathrm{Swish}(xW_1 + b_1) \otimes (xW_2 + b_2) \tag{2}$$

where x is the input tensor, W_1 and W_2 are learnable weight matrices, b_1 and b_2 are bias terms, $\mathrm{Swish}(x) = x \cdot \mathrm{sigmoid}(\beta x)$, and $\otimes$ denotes element-wise multiplication. This design enhances both computational efficiency and the compatibility between the model's visual and language components.

Text Encoder: The input question text is first processed by a tokenizer, which segments it into subword or character-level discrete units based on a predefined vocabulary, generating an ordered sequence of tokens. Furthermore, the system integrates a Knowledge Enhancement module to enhance its reasoning capabilities for traffic scenarios. The technical details and principles of this module are elaborated in Section 2.2.

Feature Fusion: The original patch features, denoted as F, undergo a four-group concatenation process to form a new feature vector F'. This vector is then projected through a two-layer Multilayer Perceptron (MLP) to match the dimensionality of the LLM's text embeddings, as described by eq.(3)

$$F'' = \sigma(F'W_1 + b_1)W_2 + b_2 \tag{3}$$

The resulting fused features are subsequently fed into the Qwen2.5 LM Decoder. This approach effectively reduces computational costs while dynamically and flexibly compressing image feature sequences of varying lengths.

2.2. Traffic-specific Knowledge Enhancement Module

To address domain adaptability limitations of general LLMs in traffic tasks, we first design a traffic-specific Knowledge Enhancement module as follows.

RAG: This study develops a traffic-oriented Retrieval-Augmented Generation (RAG) module to enhance the reasoning capacity of large language models in complex traffic scenarios. A heterogeneous corpus covering regulations, violations, abnormal events, management guidelines, and authoritative interpretations is constructed. The corpus is segmented by semantic integrity into $\mathcal{T} = t_1, t_2, \ldots, t_M$, and BERT-Base is employed to generate $d_e = 768$-dimensional embeddings, forming the database $\mathcal{D} = \boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_M$. For an input query Q, its embedding is obtained by the same encoder, $\mathbf{e}Q = \mathcal{E}\text{text}(Q; \theta_{\text{enc}})$. Semantic similarity with each chunk is computed using cosine similarity:

$$\text{sim}(Q, t_i) = \frac{\mathbf{e}_Q \cdot \mathbf{e}_i^\top}{|\mathbf{e}_Q|_2 \cdot |\mathbf{e}_i|_2 + 10^{-8}}. \tag{4}$$

All chunks are ranked accordingly, and the *Top-K* most relevant segments ($K = 5$) are selected as $t_{i_1}, \ldots, t_{i_K}$. The query and retrieved segments are concatenated into a prompt $P = [Q; t_{i_1}; \ldots; t_{i_K}]$, which serves as input to the downstream language model for knowledge-grounded reasoning in the traffic domain.

Chain-of-Thought: Experiments show that when large models handle multiple-choice reasoning tasks, directly outputting answers tends to lead to misjudgments due to incomplete reasoning chains, making it difficult to fully explore the logical connections between the information in the question stems and the options. To address this issue, we incorporate Chain-of-Thought (CoT)-guided statements into the prompt design, explicitly encouraging the model to first decompose the reasoning process step by step, and then output the final answer based on a complete reasoning chain. By forcing the model to expose its reasoning process and fill in logical gaps, this design effectively enhances the rigor of reasoning, significantly reduces the misclassification probability caused by jumpy reasoning, and improves the result reliability of multiple-choice reasoning tasks.

2.3. Training

In this section, we will introduce the training data and training methods separately.

Construction of Training Dataset: In the construction of the training data, we established a multimodal dataset for traffic event judgment and analysis. The dataset comprises DriveQA training data and a self-constructed traffic multimodal dataset. Our own dataset focuses on generative regression tasks, covering scenarios such as traffic accident determination and causal analysis, as well as abnormal driving behavior identification, with fine-grained textual labels. The DriveQA dataset concentrates on traffic signs and the associated rules from the autonomous driving perspective, with labels provided in the form of multiple-choice questions and textual explanations.

Training: For efficient large model fine-tuning, we adopt the Low-Rank Adaptation (LoRA) method, which adapts the model to the task at hand. In fine-tuning, LoRA only targets the query (Q), key (K), value (V), and output (O) projection matrices in the original transformer layers. For the weight update $\Delta W \in \mathbb{R}^{d \times d}$ of these matrices, low-rank adaptation is performed as follows: $\Delta W = BA$, where $B \in \mathbb{R}^{r \times d}$ and $A \in \mathbb{R}^{d \times r}$ (with $r \ll d$). This reduces the number of updatable parameters from d^2 to $2dr$.

During fine-tuning, all model layers are frozen except the LoRA-adapted projection matrices. Only the parameters B and A undergo gradient updates, evaluated via gradient descent as eq.(5),

$$\theta_{\text{new}} = \theta_{\text{old}} - \eta \nabla \mathcal{L}(\theta_{\text{old}}) \tag{5}$$

where η is the learning rate and $\mathcal{L}$ is the model's loss function. LoRA's parameter-efficient scheme for fine-tuning prevents original model's forgetting, reduces computational overhead, and enhances the attention mechanism's flexibility.

3. EXPERIMENT

3.1. Implementation Details

Traffic-MLLM is built upon Qwen2.5-VL with approximately 3B parameters. During training, the input video resolution is set to 640×480. In the supervised fine-tuning stage, we employ the Adam optimizer with a per-GPU batch size of 1, and the learning rate follows a linear scheduling strategy with a peak value of 10^{-5}. Thanks to the introduction of LoRA fine-tuning, GPU memory consumption is significantly reduced, enabling the training process to be completed with only four RTX3090 GPUs. During inference, the input consists of an 8-frame sequence sampled from scene videos at a resolution of 640×480. The $Top - k$ parameter in the Retrieval-Augmented Generation (RAG) module is set to 5.

3.2. Experimental Results

We presented the experimental results of the SUTD-TrafficQA and TAD datasets.

SUTD-TrafficQA: SUTD-TrafficQA is a dataset constructed from real-world traffic scenarios, containing 62,535 question-answer pairs and 10,090 videos. It consists of six challenging reasoning tasks, including basic understanding, event prediction, counterfactual reasoning, introspection, attribution analysis, and inverse reasoning. Model performance is evaluated by computing accuracy on multiple-choice questions across these tasks, with basic understanding being the most prominent and also highly challenging.

As table 1, experimental results show that our method achieves state-of-the-art overall performance on the SUTD-TrafficQA dataset, reaching an accuracy of 44.1%, significantly surpassing the previous best method CMCIR at 38.58%. Notably, in the basic understanding task, our model achieves 46.5%, substantially outperforming other methods, highlighting its superiority in semantic comprehension and fine-grained modeling for traffic scenarios. Meanwhile, in complex tasks such as attribution, counterfactual reasoning, prediction, and inverse reasoning, our method also demonstrates stable and competitive performance. These results collectively validate the effectiveness and advancement of the proposed approach across diverse traffic reasoning tasks.

DriveQA: DriveQA-V is a multimodal vision–language question-answering benchmark for evaluating VLMs' fine-grained analysis in autonomous-driving scenarios. The dataset combines CARLA-synthesized scenes with real-world annotations from Mapillary, comprising approximately 448k QA pairs and 68k images, spanning more than 220 U.S. traffic sign types (e.g., speed limits, prohibitory signs, construction warnings). The evaluation focuses on robustness

Table 1. Performance comparison on the SUTD-TrafficQA dataset across different reasoning tasks. Best results are in bold.

Method	Tasks						All
	Basic	Attribution	Introspection	Counterfactual	Forecasting	Reverse	
Random	25.00	25.00	25.00	25.00	25.00	25.00	25.00
Eclipse[12]	–	–	–	–	–	–	37.05
HCRN[14]	34.20	50.30	33.40	40.70	44.60	50.10	36.30
CMCIR[15]	36.10	**52.60**	**38.40**	46.00	**48.80**	**52.20**	38.58
Tem-Adaptor[16]	46.0	47.7	34.50	**55.0**	36.50	44.60	**46.0**
Traffic-MLLM	**46.50**	41.60	19.70	39.20	39.10	43.90	44.10

Table 2. Accuracy (%) of different MLLMs on **DriveQA-V(Signs)** across four sign categories. Best results in each column are highlighted in bold.

Model	Size	Regulatory	Warning	Guide	Temporary Control
Mini-InternVL[17]	2B	64.06	55.34	65.82	45.04
LLaVA-1.5[18]	7B	23.51	26.61	22.31	21.10
LLaVA-1.6-mistral[19]	7B	42.58	43.01	52.75	37.50
VILA-1.5[20]	8B	25.32	23.33	27.78	21.46
Traffic-MLLM	3B	**75.65**	**74.83**	**72.10**	**70.58**

real world env dataset

Table 3. Accuracy (%) on the **Real-World Mapillary** dataset under two settings: off-the-shelf and DQA-finetuned. Best performance in each column is highlighted in bold.

Model	Size	Accuracy	
		Off-The-Shelf	DQA-Finetuned
Mini-InternVL[17]	2B	57.25	68.61
LLaVA-1.5[18]	7B	40.68	52.34
LLaVA-1.6-mistral[19]	7B	53.18	57.71
VILA-1.5[20]	8B	34.38	60.86
Traffic-MLLM (Ours)	3B	**78.64**	**83.10**

Table 4. Ablation study of Qwen2.5VL-3B on TrafficQA and Real-World Mapillary datasets.

Method	DriveQA-V	Mapillary
Qwen2.5VL-3B	68.96%	71.74%
+ Finetune	72.65%	77.84
+ CoT	74.77%	80.43%
+ RAG (Ours)	**75.65%**	**83.10%**

under varied environmental conditions—lighting, weather, and camera viewpoints—and the ability to accurately recognize traffic signs and answer related regulatory questions. Notably, training data are drawn exclusively from CARLA-synthetic scenes, with no Mapillary annotations used for training, enabling assessment of real-world generalization.

Table 2 shows that Traffic-MLLM (3B) achieves leading performance across the four traffic-sign recognition tasks on the CARLA simulation dataset, with particular strength on Regulatory and Warning signs. Notably, even when compared with higher-capacity models such as LLaVA-1.5 (7B), LLaVA-1.6-mistral (7B), and VILA-1.5 (8B), Traffic-MLLM maintains a clear advantage, underscoring its superior domain adaptation and effectiveness for traffic-scene understanding.

An evaluation on the real-world Mapillary dataset (Table 3) demonstrates that Traffic-MLLM (3B) exhibits strong generalization, achieving 78.64% accuracy in an off-the-shelf setting without Mapillary fine-tuning, clearly outperforming other fine-tuned baselines. After CARLA-data fine-tuning (DQA-Finetuned), the accuracy rises to 83.10%, underscoring the model's robust performance and adaptability to multimodal traffic data

3.3. Ablation Studies

To validate the effectiveness of the proposed module in traffic video understanding tasks, this paper designs ablation experiments as follows: (1) perform inference validation using the Base model; (2) validate using the trained model; (3) enhance inference by incorporating CoT (Chain of Thought) reasoning; (4) further enhance inference by adding traffic knowledge through RAG (Retrieval-Augmented Generation).

Table 4 shows that progressively integrating the core modules yields pronounced, stepwise performance gains. The baseline Qwen model already exhibits basic traffic-scene visual understanding, achieving 68.96% on DriveQA-V and 71.74% on Mapillary. After targeted fine-tuning, the scores improve to 72.65% and 77.84%, indicating enhanced scene adaptability. Introducing a chain-of-thought (CoT) reasoning module yields a further uplift, confirming the benefit of strengthened reasoning for complex scene analysis. When combined with traffic-knowledge retrieval-augmented generation (RAG), the final results reach 75.65% on DriveQA-V and 83.10% on Mapillary. This progressive strengthening—from scene adaptation to reasoning to knowledge grounding—substantially enhances traffic video understanding and sign recognition.

4. CONCLUSION

We propose Traffic-MLLM, a domain-specific large multimodal model for traffic understanding, built on Qwen2.5-VL. By applying parameter-efficient fine-tuning (PEFT) and knowledge-prompt augmentation, we enhance multimodal data analysis capabilities. Experiments show the model achieves state-of-the-art results on DriveQA, and in a zero-shot setting outperforms several fully supervised methods in TrafficQA, demonstrating significant advantages and generalization in the traffic domain.

5. REFERENCES

[1] Noor Ul Ain Tahir, Zhe Long, Zuping Zhang, Muhammad Asim, and Mohammed ELAffendi, "Pvswin-yolov8s: Uav-based pedestrian and vehicle detection for traffic management in smart cities using improved yolov8," *Drones*, vol. 8, no. 3, pp. 84, 2024.

[2] Qiang Lu, Waikit Xiu, Xiying Li, Shenyu Hu, and Shengbo Sun, "Contrastive learning-driven traffic sign perception: Multi-modal fusion of text and vision," 2025.

[3] Ziyu Lin, Yunfan Wu, Yuhang Ma, Junzhou Chen, Ronghui Zhang, Jiaming Wu, Guodong Yin, and Liang Lin, "Yolo-llts: Real-time low-light traffic sign detection via prior-guided enhancement and multi-branch feature interaction," *IEEE Transactions on Instrumentation and Measurement*, pp. 1–1, 2025.

[4] Xusen Guo, Qiming Zhang, Junyue Jiang, Mingxing Peng, Meixin Zhu, and Hao Frank Yang, "Towards explainable traffic flow prediction with large language models," *Communications in Transportation Research*, vol. 4, pp. 100150, 2024.

[5] Peng Wang, Xiang Wei, Fangxu Hu, and Wenjuan Han, "Transgpt: Multi-modal generative pre-trained transformer for transportation," 2024.

[6] Chenxi Liu, Sun Yang, Qianxiong Xu, Zhishuai Li, Cheng Long, Ziyue Li, and Rui Zhao, "Spatial-temporal large language model for traffic prediction," in *2024 25th IEEE International Conference on Mobile Data Management (MDM)*, 2024, pp. 31–40.

[7] Lening Wang, Yilong Ren, Han Jiang, Pinlong Cai, Daocheng Fu, Tianqi Wang, Zhiyong Cui, Haiyang Yu, Xuesong Wang, Hanchu Zhou, Helai Huang, and Yinhai Wang, "Accidentgpt: Accident analysis and prevention from v2x environmental perception with multi-modal large model," 2023.

[8] 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.

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

[10] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al., "Lora: Low-rank adaptation of large language models.," *ICLR*, vol. 1, no. 2, pp. 3, 2022.

[11] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al., "Chain-of-thought prompting elicits reasoning in large language models," *Advances in neural information processing systems*, vol. 35, pp. 24824–24837, 2022.

[12] Li Xu, He Huang, and Jun Liu, "Sutd-trafficqa: A question answering benchmark and an efficient network for video reasoning over traffic events," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 9878–9888.

[13] Maolin Wei, Wanzhou Liu, and Eshed Ohn-Bar, "Passing the driving knowledge test," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025.

[14] Thao Minh Le, Vuong Le, Svetha Venkatesh, and Truyen Tran, "Hierarchical conditional relation networks for video question answering," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 9972–9981.

[15] Yang Liu, Guanbin Li, and Liang Lin, "Cross-modal causal relational reasoning for event-level visual question answering," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 10, pp. 11624–11641, 2023.

[16] Guangyi Chen, Xiao Liu, Guangrun Wang, Kun Zhang, Philip HS Torr, Xiao-Ping Zhang, and Yansong Tang, "Tem-adapter: Adapting image-text pretraining for video question answer," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 13945–13955.

[17] Zhangwei Gao, Zhe Chen, Erfei Cui, Yiming Ren, Weiyun Wang, Jinguo Zhu, Hao Tian, Shenglong Ye, Junjun He, Xizhou Zhu, et al., "Mini-internvl: a flexible-transfer pocket multi-modal model with 5% parameters and 90% performance," *Visual Intelligence*, vol. 2, no. 1, pp. 32, 2024.

[18] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee, "Improved baselines with visual instruction tuning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2024, pp. 26296–26306.

[19] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee, "Llavanext: Improved reasoning, ocr, and world knowledge," 2024.

[20] Ji Lin, Hongxu Yin, Wei Ping, Pavlo Molchanov, Mohammad Shoeybi, and Song Han, "Vila: On pre-training for visual language models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2024, pp. 26689–26699.