

# Task-Specific Dual-Model Framework for Comprehensive Traffic Safety Video Description and Analysis

Blessing Agyei Kyem

North Dakota State University

Fargo, North Dakota, North Dakota

blessing.agyeikyem@ndsu.edu

Eugene Denteh

North Dakota State University

Fargo, North Dakota, North Dakota

eugene.denteh@ndsu.edu

Andrews Danyo

North Dakota State University

Fargo, North Dakota, North Dakota

andrews.danyo@ndsu.edu

## Abstract

Traffic safety analysis requires complex video understanding to capture fine-grained behavioral patterns and generate comprehensive descriptions for accident prevention. In this work, we present a unique dual-model framework that strategically utilizes the complementary strengths of VideoLLaMA and Qwen2.5-VL through task-specific optimization to address this issue. The core insight behind our approach is that separating training for captioning and visual question answering (VQA) tasks minimizes task interference and allows each model to specialize more effectively. Experimental results demonstrate that VideoLLaMA is particularly effective in temporal reasoning, achieving a CIDEr score of 1.1001, while Qwen2.5-VL excels in visual understanding with a VQA accuracy of 60.80%. Through extensive experiments on the WTS dataset, our method achieves an S2 score of 45.7572. Ablation studies validate that our separate training strategy outperforms joint training by 8.6% in VQA accuracy while maintaining captioning quality.

## Keywords

Traffic Safety Analysis, Multi-modal LLMs, Vision Language Models, Task Interference, Video Captioning, Visual Question Answering

## ACM Reference Format:

Blessing Agyei Kyem, Eugene Denteh, and Andrews Danyo. 2025. Task-Specific Dual-Model Framework for Comprehensive Traffic Safety Video Description and Analysis. In . ACM, New York, NY, USA, 9 pages. <https://doi.org/10.1145/nmnnnn.nmnnnn>

## 1 Introduction

Fine-grained video captioning of traffic safety scenarios has emerged as a transformative approach for understanding and preventing pedestrian accidents, enabling detailed analysis of behavioral patterns, environmental conditions, and contextual factors that traditional detection systems fail to capture [17, 27]. The complexity of generating comprehensive textual descriptions from multi-perspective traffic footage presents unique challenges in computer vision, requiring advanced techniques that can synthesize temporal dynamics, spatial relationships, and causal narratives across continuous video sequences [16]. Currently deployed ITS predominantly

---

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](mailto:permissions@acm.org).

*Conference'17, Washington, DC, USA*

© 2025 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM

<https://doi.org/10.1145/nmnnnn.nmnnnn>

rely on traditional post-incident reconstruction methods and binary classification systems, offering limited descriptive capabilities in complex pedestrian accident scenarios [29, 47]. Comprehensive analysis systems require enhanced coverage and a detailed understanding of behavioral patterns to overcome these limitations [45]. However, these conventional approaches often fail to deliver continuous and detailed representations of the monitored scenarios, which remains a fundamental obstacle in developing proactive safety management systems.

Advanced video understanding now employs long fine-grained captioning to analyze traffic behaviors, particularly in vehicle-pedestrian accident scenarios [13, 15] (see Fig. 2). This method facilitates continuous monitoring and detailed description of complex traffic events throughout their temporal progression using synchronized footage from multiple cameras and viewpoints, including static overhead cameras and moving vehicle ego cameras [49] as illustrated in Fig. 2. This includes TrafficVLM [13] which analyzes traffic video events at various spatial and temporal resolutions, generating extended, detailed descriptions covering both vehicle and pedestrian activities across different phases of an incident [40, 69]. This enhanced analytical capability significantly advances traditional binary classification systems by providing rich temporal narratives that capture behavioral patterns, environmental conditions, and evolving risk factors integral to incident development [3, 35].

Current video captioning methods face critical limitations in multi-perspective traffic safety scenarios [55]. VideoLLaMA [61] and Video-LLaVA [30] excel at long context understanding and temporal modeling, generating comprehensive captions across extended sequences [62], but suffer from hallucination artifacts and data efficiency constraints [34]. Conversely, Qwen-VL [7] and InternVL [11] demonstrate superior visual reasoning and state-of-the-art VQA performance through robust spatial understanding [8], yet struggle with extended temporal contexts and computational scalability [53]. Both architectures face synchronization overhead in multi-camera deployments, limiting real-time applications. Integrating Video-LLaMA's temporal reasoning with Qwen-VL's visual reasoning presents a complementary solution where each model's strengths address the other's limitations.

To address these challenges, we propose a novel approach combining VideoLLaMA and QwenVL for traffic safety analysis. Our method utilizes VideoLLaMA's temporal reasoning and narrative generation with QwenVL's visual reasoning and spatial analysis. Our contributions have been summarized below:

- We propose an approach that systematically integrates VideoLLaMA's temporal reasoning with Qwen-VL's visual reasoning capabilities for traffic safety analysis.

- We empirically validate that separate task-specific training for captioning and visual question answering significantly outperforms joint training approaches.
- We conduct extensive experiments including ablation studies on the WTS dataset across multiple metrics.

## 2 Related Work

### 2.1 Video Captioning and Multimodal Analysis in Traffic Safety

Early traffic safety studies [18, 46] relied on manual observation providing foundational insights but limited by subjectivity and scale [37]. To address these limitations, Haritaoglu et al. [20] advanced the field with the W4 surveillance system for automated tracking, though single-camera constraints still restricted comprehensive analysis. Building on automated approaches, Krishna et al. [26] introduced dense-captioning methods for simultaneous event descriptions, which Shoman et al. [50] subsequently adapted specifically for traffic safety by integrating visual-textual features. These captioning capabilities were further enhanced when Zhang et al. [64] employed language models for generating narrative descriptions from complex visual data. Supporting this research direction, KongKong et al. [25] released the WTS dataset, providing essential resources for detailed pedestrian-vehicle analysis [1, 23].

### 2.2 Pedestrian-Vehicle Interaction and Accident Prediction

A fundamental challenge in traffic safety research has been integrating fragmented multi-camera data into unified spatiotemporal representations [5, 6, 52]. While early multi-camera efforts [14] struggled with behavioral coherence, transformer-based approaches by Zhou et al. [68] successfully integrated long-term patterns across multiple feeds [39, 42], setting new standards for traffic analysis [57]. Wu et al. [58] further advanced this by synthesizing multimodal camera data for improved accident prediction. Concurrently, trajectory prediction evolved from Alahi et al.'s [2] Social-LSTM, which suffered from pixel-based limitations, to attention-enhanced models [32] and graph-based approaches [4] that better capture spatiotemporal interactions. However, Behboudi et al. [9] note that real-time performance remains a critical challenge, underscoring the need for continued methodological advancement.

### 2.3 Vision Language Models for Traffic Captioning and Reasoning

The past two years have seen specialized VLMs such as TrafficVLM [13], Wolf [28] and VLM-Auto [19] set new standards in traffic video captioning by combining multi-task fine-tuning, mixture-of-experts backbones and simulator-integrated pipelines. At the same time, general models such as VideoLLaMA [12, 61, 63], InternVL [11] and Qwen-VL [7] have been effectively adapted via domain fine-tuning and LoRA to achieve up to 60% VQA accuracy in pedestrian-vehicle reasoning [67]. Dedicated VQA benchmarks including NuScenes-QA [43], LingoQA [36] and DriveLM [51] have further advanced traffic-specific visual reasoning through graph-based and free-form QA formats, exposing gaps such as GPT-4V's 59.6 % truthfulness score [36, 38, 41].

### 2.4 Architectural Advances, Benchmarks, and Deployment Challenges

Domain-specialized architectures such as EMMA [22], DriveVLM [54] and MAPLM [10] utilize end-to-end multimodal fusion, chain-of-thought planning and HD-map integration to tackle safety-critical reasoning and spatiotemporal modeling. Hybrid training strategies such as joint multi task learning [66], supervised with contrastive learning [44, 59], and LoRA fine tuning [21] consistently outperform separate pipelines [48]. Benchmarks like WTS [25] and the AI City Challenge competition [56] offer rigorous evaluation of caption quality, VQA accuracy, and event phase understanding. However, persistent challenges remain in hallucination mitigation, balancing spatial resolution against temporal coverage, and achieving real-time deployment on edge hardware prompting solutions like Visual-Textual Intervention and encoder ensembles [33].

## 3 Proposed Method

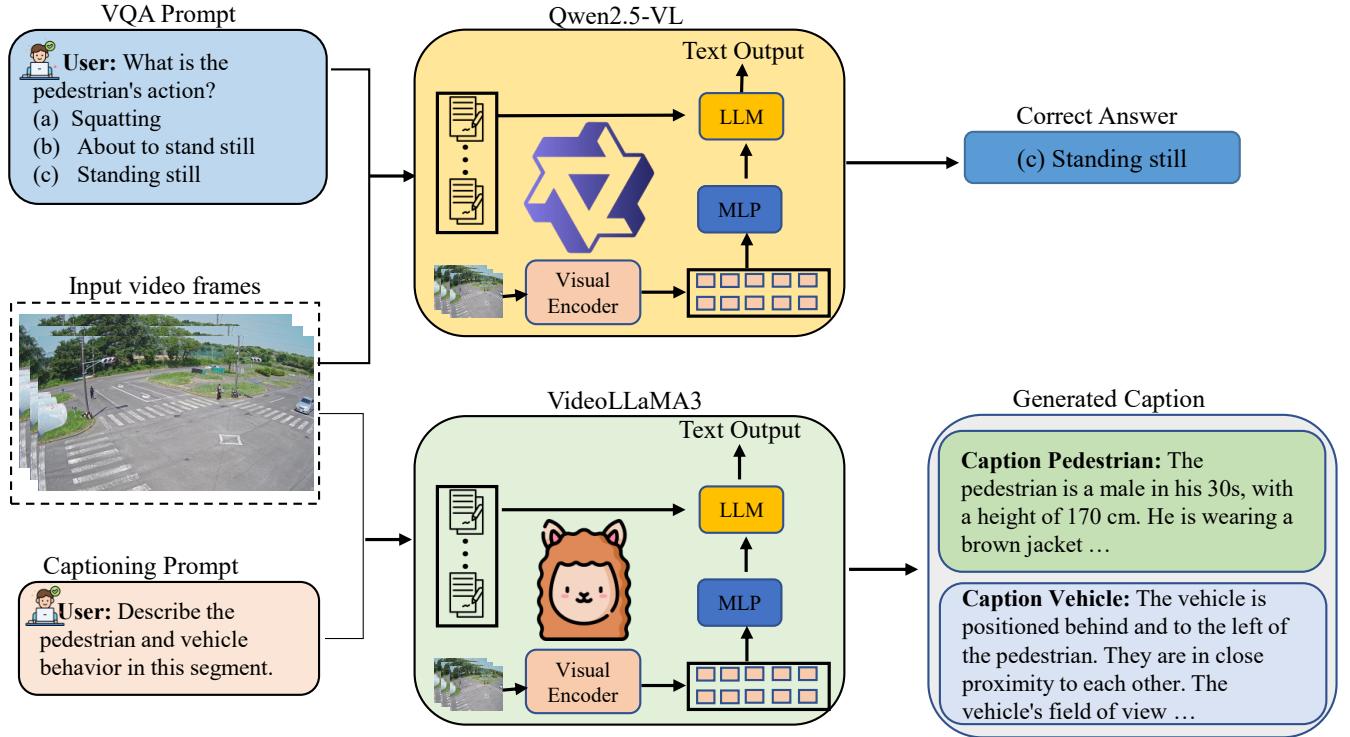
Figure 1 illustrates our proposed dual-task optimisation framework, which uses two specialized models for comprehensive traffic safety analysis. The framework processes input video frames through two parallel pathways: Qwen2.5-VL handles Visual Question Answering (VQA) tasks. Each model is independently optimized for its specific task using Low-Rank Adaptation (LoRA), preventing task interference while maximizing individual model performance.

### 3.1 Training Pipeline

Our training pipeline involves two primary tasks: fine-grained video captioning and visual question answering (VQA). Both tasks utilize segments extracted from videos in the WTS dataset, each associated with precise timestamps and behavior phases.

**Captioning Task** We begin by extracting video segments based on annotated timestamps corresponding to distinct behavioral phases. Each video segment is paired with detailed captions that describe the behaviors of pedestrians and vehicles, consistent with the structured annotation approach used in the WTS dataset. These segment-caption pairs are then input into our chosen Video-Large Language Models (Video-LLMs), specifically QwenVL-2.5 and VideoLLaMA, for instruction fine-tuning using Low-Rank Adaptation (LoRA). The selection of QwenVL-2.5 and VideoLLaMA is motivated by their superior multimodal understanding capabilities, efficient handling of fine-grained temporal and spatial information, and their demonstrated state-of-the-art performance in complex multimodal benchmarks. For inference, we apply the same segment extraction process. The fine-tuned Video-LLMs generate continuous and detailed captions for pedestrian and vehicle behaviours within these segments, thereby providing context-rich descriptions closely aligned with the ground truth annotations.

**Visual Question Answering Task** We follow a similar process for segment extraction. However, the segments are paired with questions and predefined multiple-choice answers, which are extracted directly from the annotated dataset. The Video-LLMs undergo separate instruction fine-tuning with this question-answer data, adapting specifically to VQA scenarios. During VQA inference, we again extract relevant video segments and provide them to the fine-tuned Video-LLMs alongside corresponding questions. Given



**Figure 1: Overview of our proposed dual-task optimization framework.**

the multiple-choice nature of the task, we employ greedy decoding during inference to select the most likely answer option. This approach ensures precise, singular answers suitable for objective evaluation.

### 3.2 Problem Formulation

Given a video  $\mathcal{V} = \{v_t\}_{t=1}^T$  with  $T$  frames segmented into  $N$  behavioral phases  $\mathcal{S} = \{s_i\}_{i=1}^N$ , we address two complementary tasks: (1) generating fine-grained captions  $C = \{c_p^i, c_v^i\}_{i=1}^N$  describing pedestrian ( $p$ ) and vehicle ( $v$ ) behaviors, and (2) answering  $M$  safety-critical questions  $Q = \{q_j\}_{j=1}^M$  with corresponding answers  $\mathcal{A} = \{a_j\}_{j=1}^M$ .

### 3.3 Dual-Task Optimization Framework

We employ VideoLLaMA3 with parameters  $\theta$  (denoted as  $f_\theta$ ) and Qwen2.5-VL with parameters  $\phi$  (denoted as  $g_\phi$ ) enhanced with task-specific Low-Rank Adaptation (LoRA). For each model's weight matrix  $W \in \mathbb{R}^{d \times d}$ , we define separate LoRA decompositions:

$$W' = W + \Delta W, \quad (1)$$

$$\text{where } \Delta W = BA, \quad B \in \mathbb{R}^{d \times r}, A \in \mathbb{R}^{r \times d}$$

with rank  $r \ll d$ , where  $d$  is the model dimension. We optimize distinct parameters  $\{B_{cap}, A_{cap}\}$  for captioning and  $\{B_{vqa}, A_{vqa}\}$  for VQA.

**Captioning Objective.** For each segment  $s_i$ , we minimize the negative log-likelihood of generating captions:

$$\mathcal{L}_{cap} = - \sum_{i=1}^N \sum_{k \in \{p, v\}} \sum_{t=1}^{L_k^i} \log P(c_k^{i,t} | c_k^{i,<t}, s_i; \theta, B_{cap}, A_{cap}) \quad (2)$$

where  $c_k^{i,t}$  denotes the  $t$ -th token of caption type  $k \in \{p, v\}$  for segment  $i$ ,  $c_k^{i,<t} = \{c_k^{i,1}, \dots, c_k^{i,t-1}\}$  represents preceding tokens,  $L_k^i$  is the length of caption  $k$  for segment  $i$ , and  $P$  denotes the probability distribution over the vocabulary.

**VQA Objective.** Given question-answer pairs, we optimize:

$$\mathcal{L}_{vqa} = - \sum_{j=1}^M \log P(a_j | q_j, s_j; \phi, B_{vqa}, A_{vqa}) \quad (3)$$

where  $a_j \in \{1, \dots, K\}$  represents the correct answer among  $K$  multiple-choice options for question  $q_j$  on segment  $s_j$ .

### 3.4 Training Strategy

We adopt separate training to prevent task interference:

$$\begin{aligned} \{\theta^*, B_{cap}^*, A_{cap}^*\} &= \arg \min_{\theta, B_{cap}, A_{cap}} \mathcal{L}_{cap} \\ \{\phi^*, B_{vqa}^*, A_{vqa}^*\} &= \arg \min_{\phi, B_{vqa}, A_{vqa}} \mathcal{L}_{vqa} \end{aligned} \quad (4)$$

where  $*$  denotes optimal parameters. This decoupled optimization preserves each model's specialized capabilities.

### 3.5 Inference

During inference, for a test segment  $s_{test}$ :

**Caption Generation:** We employ beam search with beam width  $b$  to generate caption  $\hat{c}_k$ :

$$\hat{c}_k = \arg \max_c P(c|s_{test}; \theta^*, B_{cap}^*, A_{cap}^*)$$

**VQA:** For question  $q$ , we select the predicted answer  $\hat{a}$ :

$$\hat{a} = \arg \max_{a \in \{1, \dots, K\}} P(a|q, s_{test}; \phi^*, B_{vqa}^*, A_{vqa}^*) \quad (5)$$

The complementary outputs provide a comprehensive understanding of the traffic scene.

## 4 Experiment

### 4.1 Datasets

Our model is trained and evaluated using the WTS dataset [24] developed by Toyota, Inc. The dataset contains 810 internal videos (155 scenarios) and 3,400 external pedestrian-related clips from BDD100K, representing over 130 staged traffic scenarios captured at 1080p/30fps using vehicle-mounted and overhead cameras. Each video is segmented into five behavioral phases (pre-recognition, recognition, judgment, action, avoidance) with dual-perspective captions (pedestrian and vehicle viewpoints). The dataset includes manually annotated bounding boxes with object tracking, 3D pedestrian gaze vectors, and a Traffic VQA dataset with 180 structured questions for spatial-temporal understanding. Figure 2 shows the structure of WTS dataset.

### 4.2 Evaluation Metrics

The evaluation framework for our approach employs a dual-task assessment methodology that comprehensively measures both natural language generation quality and visual reasoning capabilities through two complementary sub-tasks.

**4.2.1 Caption Generation.** The caption generation component evaluates the quality of automatically generated natural language descriptions for traffic safety scenarios. For each video segment, models must produce two detailed captions: one describing pedestrian behavior and one describing vehicle behavior throughout the temporal sequence. Performance is assessed using a composite metric that averages four established natural language generation measures: BLEU-4, METEOR, ROUGE-L, and CIDEr.

**BLEU-4** evaluates n-gram precision between generated and reference captions, defined as:

$$\text{BLEU-4} = BP \cdot \exp \left( \sum_{n=1}^4 w_n \log p_n \right) \quad (6)$$

where  $p_n$  is the modified n-gram precision,  $w_n = 1/4$  for uniform weighting, and  $BP$  is the brevity penalty to discourage overly short generations.

**METEOR** evaluates translation quality by combining precision-recall harmonic mean with a penalty term that accounts for word order fragmentation:

$$\text{METEOR} = F_{\text{mean}} \times (1 - \text{Penalty}) \quad (7)$$

The harmonic mean of precision and recall is computed as:

$$F_{\text{mean}} = \frac{P \cdot R}{\alpha \cdot P + (1 - \alpha) \cdot R} \quad (8)$$

where precision  $P$  and recall  $R$  are defined in terms of matched unigrams:

$$P = \frac{m}{w_t}, \quad R = \frac{m}{w_r} \quad (9)$$

The fragmentation penalty is calculated based on the number of contiguous chunks:

$$\text{Penalty} = \gamma \cdot \left( \frac{ch}{m} \right)^\beta \quad (10)$$

Here,  $m$  represents matched unigrams between hypothesis and reference,  $w_t$  and  $w_r$  denote the total unigrams in hypothesis and reference respectively,  $ch$  indicates the number of contiguous matched chunks, with standard parameter values  $\alpha = 0.9$ ,  $\beta = 3$ , and  $\gamma = 0.5$ .

**ROUGE-L** evaluates caption quality by measuring sequence-level similarity through longest common subsequences, which is particularly relevant for traffic safety descriptions where maintaining temporal coherence is crucial. The metric is formulated as

$$\text{ROUGE-L} = \frac{(1 + \beta^2) \times R_{lcs} \times P_{lcs}}{R_{lcs} + \beta^2 \times P_{lcs}} \quad (11)$$

where:

$$R_{lcs} = \frac{\text{LCS}(X, Y)}{m}, \quad P_{lcs} = \frac{\text{LCS}(X, Y)}{n} \quad (12)$$

where  $\text{LCS}(X, Y)$  is the length of longest common subsequence between candidate  $X$  and reference  $Y$ ,  $m$  is the length of reference sequence,  $n$  is the length of candidate sequence, and  $\beta = P_{lcs}/R_{lcs}$  when  $R_{lcs} \neq 0$ .

**CIDEr** measures the consensus between generated captions and human references by emphasizing n-grams that are distinctive and informative, making it suitable for evaluating detailed traffic safety descriptions that require specific terminology and behavioral descriptors.

$$\text{CIDEr}_n(c_i, S_i) = \frac{1}{m} \sum_j \frac{g^n(c_i) \cdot g^n(s_{ij})}{|g^n(c_i)| |g^n(s_{ij})|} \quad (13)$$

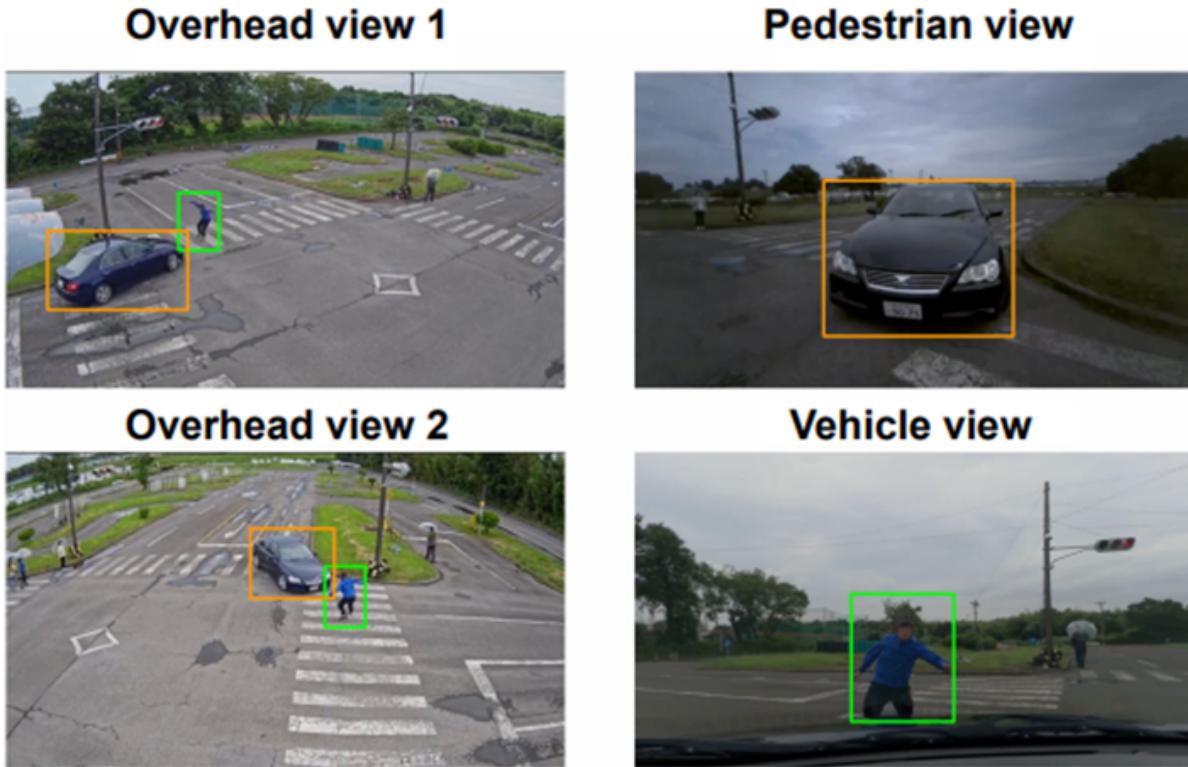
where  $g_n(s)$  represents the weighted word combinations for caption  $s$ ,  $m$  is the number of reference captions, and  $N$  is the maximum word combination length.

**4.2.2 Visual Question Answering.** The second sub-task measures the model's capability to reason visually and provide accurate answers to safety-critical questions. The performance metric employed is simple top-1 accuracy:

$$\text{Acc} = \frac{1}{N} \sum_{i=1}^N \mathbf{1}[\hat{a}_i = a_i^*] \quad (14)$$

where  $N$  is the total number of questions,  $\hat{a}_i$  is the predicted answer,  $a_i^*$  the ground-truth answer, and  $\mathbf{1}[\cdot]$  denotes the indicator function.

The overall performance is summarized into a composite final metric ( $S_2$ ), calculated by combining the caption generation (Cap\_Score) and VQA accuracy (Acc) metrics as follows:



**Caption pedestrian:** The pedestrian, a male in his 20s and about 170 cm tall, was walking across a zebra-striped crosswalk situated at the centre of a four-way urban intersection. He wore a vivid blue hooded jacket and dark trousers. The weather was overcast yet bright, with dry, level asphalt and no standing water. His body was oriented perpendicular to the vehicle's path, moving from the driver's right toward the opposite curb. Although he cast a brief glance toward the roadway, he seemed mainly focused on reaching the other side, showing little reaction to the approaching sedan. The relative distance between pedestrian and vehicle was moderate and decreasing, while surrounding traffic remained light. It was a weekday, and the pedestrian seemed to be aware of the approaching vehicle.

**Caption vehicle:** The vehicle was moving at a constant speed of 10km/h. It trailed the pedestrian by few metres and remained diagonal to the crosswalk, slightly to the pedestrian's left. The vehicle had a clear view of the pedestrian. It was going diagonally and towards the direction of the pedestrian. The environment conditions indicated that the pedestrian was a male in his 20s with a height of 170 cm. He was wearing a blue hooded jacket and dark trousers. The event took place in an urban area on a weekday. The weather was overcast, but the brightness was bright. The road surface was dry and level, made of asphalt. The traffic volume was very low on all the lanes on the road. Sidewalks were present on both sides of the road.

**Figure 2: Multi-view sample from the WTS caption dataset: Overhead Views 1 & 2 (left), Pedestrian View (top-right), and Vehicle View (bottom-right).**

$$\text{Cap\_Score} = \frac{\text{BLEU-4} + \text{METEOR} + \text{ROUGE-L} + \text{CIDEr}}{4} \quad (15)$$

$$S2 = \frac{\text{Cap\_Score} + \text{Acc}}{2} \quad (16)$$

### 4.3 Implementation Details

We implement our experiments using two state-of-the-art Video-Large Language Models: Qwen2.5-VL and VideoLLaMA3.

**Qwen2.5-VL.** Qwen2.5-VL integrates a dynamic-resolution Vision Transformer backbone that processes frames at varying sizes without costly normalization, paired with a specialized multimodal projector to align visual tokens with the language embedding space [7]. It achieves precise object grounding via both bounding-box and point-based outputs, and handles long videos using absolute time encodings and dynamic frame sampling. We train this model with a batch size of 4, gradient accumulation over 8 steps for 10 epochs at a learning rate of  $1 \times 10^{-4}$ , using a cosine scheduler with a 0.1 warm-up ratio.

**VideoLLaMA3.** VideoLLaMA3 couples a pretrained vision encoder and a LLaMA-based language model through a lightweight adaptor, accepting variable-resolution inputs and merging similar frames at inference to reduce token counts [61]. Its core joint vision–language alignment module is first co-trained on large-scale image-text corpora, then fine-tuned in a multi-task stage with both image-text and video-text supervision, followed by a video-centric refinement that enhances temporal consistency and motion reasoning. Training uses learning rates of  $1 \times 10^{-5}$  for the LLM and projector and  $2 \times 10^{-6}$  for the vision encoder, weight decay 0.01, gradient accumulation of 4 steps, gradient checkpointing, and a cosine scheduler with a 0.03 warm-up ratio, logging every 5 steps. This complementary approach combines Qwen2.5-VL's precise visual understanding with VideoLLaMA3's coherent temporal narrative generation to provide a comprehensive analysis of traffic scenes.

## 5 Results and Discussion

Table 1 presents our experimental results comparing the proposed dual-model approach against state-of-the-art baselines. Our integrated VideoLLaMA3-7B and Qwen2.5-VL-7B framework achieves an S2 score of 45.7572, significantly outperforming all individual baselines including LLaVA-NeXT-Video-7B (40.38), InternVL3-8B (41.98), and the best single model VideoLLaMA3-7B (44.73). This 1.09-point improvement over the strongest baseline validates our architectural complementarity hypothesis.

The performance gain stems from utilizing each model's specialized strengths. VideoLLaMA3-7B demonstrates superior captioning with the highest scores across all metrics (BLEU-4: 0.2569, METEOR: 0.4528, ROUGE-L: 0.4512, CIDEr: 1.1001), while Qwen2.5-VL-7B excels at visual reasoning with 60.80% VQA accuracy. Individual baselines struggle to excel at both tasks simultaneously. For example, LLaVA-NeXT-Video performs well in captioning but achieves only 49.32% VQA accuracy, while InternVL3-8B shows strong VQA performance (59.93%) but weaker captioning metrics.

Figure 3 illustrates why this complementarity matters. In a safety-critical scenario where a pedestrian uses a smartphone while standing still, VideoLLaMA3-7B accurately captures visual details (black T-shirt, stationary behavior) and generates coherent narratives. Conversely, Qwen2.5-VL produces significant hallucinations, misidentifying clothing as "brown jacket and green slacks" and behavior as "rushing out at high speed." These qualitative differences demonstrate that VideoLLaMA3-7B provides essential visual grounding inherent in its architecture for safety descriptions, while Qwen2.5-VL contributes structured visual reasoning capabilities, creating a robust approach that surpasses individual model limitations.

### 5.1 Ablation Study

We conduct an ablation study to validate our training strategy by examining the effects of training on captioning data alone, combined captioning and VQA datasets, and our proposed approach of training separately on captioning and VQA tasks. Table 2 summarizes our findings clearly.

**Training Only on Captioning Dataset.** When models are trained only on captioning data, we observe high captioning metrics but notably reduced VQA accuracy. For example, VideoLLaMA3-7B attains a CIDEr score of 1.0625 but a low VQA accuracy of 47.50%. Qwen2.5-VL-7B similarly exhibits CIDEr: 0.5500 and VQA accuracy: 49.00%. This indicates exclusive caption training limits visual reasoning capabilities.

**Training on Combined Captioning and VQA Datasets.** Joint training moderately improves visual reasoning but slightly reduces captioning quality due to task interference. VideoLLaMA3-7B achieves CIDEr: 1.0350 and VQA accuracy: 53.50%. Likewise, Qwen2.5-VL-7B demonstrates improved visual reasoning (VQA accuracy: 57.50%) with reduced captioning performance (CIDEr: 0.5450).

**Training on Separate Captioning and VQA Datasets (Our Approach).** Our proposed approach which involves training separately on captioning and VQA datasets clearly outperforms other strategies. VideoLLaMA3-7B achieves optimal captioning scores (CIDEr: 1.1001) and strong VQA accuracy (58.6121%). Qwen2.5-VL-7B records the highest VQA accuracy (60.7980%) alongside good

captioning performance (CIDEr: 0.5769). This strategy effectively addresses task interference and enhances and complementary interplay performance.

**Table 2: Ablation results highlighting the training strategies on captioning and VQA tasks.**

Training Method	Model	BLEU-4↑	CIDEr↑	VQA Acc↑
Captioning Only	VideoLLaMA3-7B	0.2500	1.0625	47.50
	Qwen2.5-VL-7B	0.1850	0.5500	49.00
Combined Tasks	VideoLLaMA3-7B	0.2420	1.0350	53.50
	Qwen2.5-VL-7B	0.1830	0.5450	57.50
Separate Tasks (Ours)	VideoLLaMA3-7B	<b>0.2569</b>	<b>1.1001</b>	<b>58.6121</b>
	Qwen2.5-VL-7B	0.1921	0.5769	60.7980

## 6 Conclusion

We presented a dual-model framework that utilizes the complementary strengths of VideoLLaMA3-7B and Qwen2.5-VL-7B for traffic safety analysis. By employing task-specific training, our approach achieves an S2 score of 45.7572, outperforming single-model baselines by 1.09 points. This improvement stems from combining VideoLLaMA's temporal reasoning capabilities with Qwen-VL's superior visual understanding, addressing the inherent trade-off between caption generation and VQA accuracy. The success of our method demonstrates that architectural complementarity, when properly optimized, can overcome the limitations of individual models in safety-critical applications. Future work will explore model distillation to maintain this performance while reducing computational requirements for real-world deployment.

## References

- [1] Asma Abdallah and Ahmed Kaabi. 2024. *Cutting-Edge Innovations in Teaching, Leadership, Technology, and Assessment*. doi:10.4018/979-8-3693-0880-6
- [2] Alexandre Alahi, Kratarth Goel, Vignesh Ramanathan, Alexandre Robicquet, Li Fei-Fei, and Silvio Savarese. 2016. Social LSTM: Human trajectory prediction in crowded spaces. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 961–971. doi:10.1109/CVPR.2016.110
- [3] A.S. Albahri, Yahya Layth Khaleel, Mustafa Abdulfattah Habeeb, Reem D. Ismael, Qabas A. Hameed, Muhammet Devci, Raad Z. Homod, O.S. Albahri, A.H. Alamoodi, and Laith Alzubaidi. 2024. A systematic review of trustworthy artificial intelligence applications in natural disasters. *Computers and Electrical Engineering* 118 (2024), 109409. doi:10.1016/j.compeleceng.2024.109409
- [4] Hameed Alghodhaifi and Sridhar Lakshmanan. 2023. Holistic Spatio-Temporal Graph Attention for Trajectory Prediction in Vehicle-Pedestrian Interactions. *Sensors* 23, 17 (2023), 7361. doi:10.3390/s23177361
- [5] Temitope Ibrahim Amosa, Patrick Sebastian, Lila Iznita Izhar, Oladimeji Ibrahim, Lukman Shehu Ayinla, Abdullah Bahashwan, Abubakar Bala, and Yan Alhaji Samaila. 2023. Multi-camera multi-object tracking: A review of current trends and future advances. *Neurocomputing* 552 (2023), 126558. doi:10.1016/j.neucom.2023.126558
- [6] Jeongyeop Baek and Lisa Lim. 2024. Unveiling pedestrian injury risk factors through integration of urban contexts using multimodal deep learning. *Sustainable Cities and Society* 101 (2024), 105168. doi:10.1016/j.scs.2023.105168
- [7] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-VL: A Versatile Vision-Language Model for Understanding, Localization, Text Reading, and Beyond. <https://api.semanticscholar.org/CorpusID:261101015>
- [8] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-VL: A Versatile Vision-Language Model for Understanding, Localization, Text Reading, and Beyond. arXiv:2308.12966 [cs.CV] <https://arxiv.org/abs/2308.12966>
- [9] Noushin Behboudi, Sobhan Moosavi, and Rajiv Rammath. 2024. Recent Advances in Traffic Accident Analysis and Prediction: A Comprehensive Review of Machine Learning Techniques. *arXiv preprint arXiv:2406.13968* (2024). arXiv:2406.13968

**Table 1: Comparison of baselines and ours on video captioning and VQA. Subscripts  $i$  and  $e$  denote the metrics computed on the internal and external splits of the WTS dataset, respectively.**

Model	Captioning Metrics								VQA Metric	Overall Metric
	BLEU-4 $_i$ ↑	METEOR $_i$ ↑	ROUGE-L $_i$ ↑	CIDEr $_i$ ↑	BLEU-4 $_e$ ↑	METEOR $_e$ ↑	ROUGE-L $_e$ ↑	CIDEr $_e$ ↑		
Acc↑	S2									
VideoLLaMA3-7B [61]	0.2569	0.4528	0.4512	1.1001	0.2814	0.4844	0.4658	1.2579	58.6121	44.7329
Qwen2.5-VL-7B [7]	0.1921	0.3968	0.3915	0.5769	0.1475	0.3483	0.3414	0.6304	60.7980	42.5136
LLaVA-NeXT-Video-7B [65]	0.2412	0.4458	0.4415	1.0122	0.2603	0.4683	0.4441	1.1276	49.3216	40.3807
Video-LLaVA-7B [31]	0.1780	0.3553	0.3548	0.7606	0.1642	0.3201	0.3184	0.7721	47.0146	35.0327
MiniCPM-V-2_6 [60]	0.2290	0.4417	0.4352	0.8504	0.2458	0.4534	0.4429	0.9656	51.3412	40.8556
InternVL3-8B [11]	0.1918	0.3951	0.3812	0.5753	0.1469	0.3475	0.3399	0.6237	59.9310	41.9799
<b>VideoLLaMA3-7B &amp; Qwen2.5-VL-7B (Ours)</b>	<b>0.2569</b>	<b>0.4528</b>	<b>0.4512</b>	<b>1.1001</b>	<b>0.2814</b>	<b>0.4844</b>	<b>0.4658</b>	<b>1.2579</b>	<b>60.7980</b>	<b>45.7572</b>



☒ **Prompt:** Describe the pedestrian and vehicle behavior in this segment.

ⓘ **Ground Truth:** The pedestrian, a male in his 20s, stood still on a suburban residential road. He was perpendicular to the vehicle and positioned diagonally to the right, in front of it. Despite the proximity, the pedestrian was unaware of the vehicle's presence as his line of sight was fixated on his smartphone held in his hand. The pedestrian's clothing consisted of a black T-shirt on the upper body and navy blue slacks on the lower body. With a height of 170 cm, he stood tall on the level asphalt road surface. The weather was clear, with bright lighting and dry road conditions. The road was quiet, belonging to a light traffic volume residential area, with two lanes for two-way traffic. Sidewalks were absent on both sides, as well as roadside strips. In this particular situation, the pedestrian remained stationary, seemingly engrossed in his smartphone, completely unaware of the vehicle and its surroundings.

ⓘ **VideoLLaMA3-7B:** The pedestrian, a male in his 30s, stood still on a residential road that was brightly lit by street lights. Wearing a black T-shirt and slacks, he had a close distance from the vehicle that approached him. His line of sight was in front, following the direction of travel, and he closely watched the vehicle. Despite being unaware of the vehicle's presence, the pedestrian noticed it and knew about its existence. The environment conditions were clear with a dry road surface, and the road had two-way traffic with light volume. There were no sidewalks or roadside strips on both sides, but there were street lights illuminating the surroundings. The pedestrian's visual status indicated that he could clearly see his surroundings. In this specific instance, the pedestrian and the vehicle were both stationary, suggesting a static and potentially calm event.

ⓘ **Qwen2.5-VL:** The pedestrian is a male in his 30s, with a height of 170 cm. He is wearing a brown jacket on his upper body and green slacks on his lower body. His orientation is perpendicular to the vehicle and to the right. He is positioned diagonally to the left in front of the vehicle, at a near distance. The pedestrian's line of sight is in front, in the direction of travel. He is closely watching the vehicle, but he is unaware of its presence. The pedestrian's general action is to go straight ahead, but he is rushing out at high speed. The environment conditions are clear with bright brightness. The road surface is dry and inclined uphill. It is made of asphalt and is a residential road with two-way traffic. The sidewalk is on both sides.

**Figure 3: Qualitative comparison of model outputs for traffic safety description. Given input video frames of a pedestrian-vehicle scenario, we compare ground truth annotations with predictions from VideoLLaMA3-7B and Qwen2.5-VL.**

- [10] Xu Cao, Tong Zhou, Yunsheng Ma, Wenqian Ye, Can Cui, Kun Tang, Zhipeng Cao, Kaizhao Liang, Ziran Wang, James M Rehg, and Chao Zheng. 2024. MAPLM: A Real-World Large-Scale Vision-Language Benchmark for Map and Traffic Scene Understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 21819–21830.
- [11] Zhe Chen, Jiannan Wu, Wenhui Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. 2024. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 24185–24198.
- [12] Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, and Lidong Bing. 2024. VideoLLaMA 2: Advancing Spatial-Temporal Modeling and Audio Understanding in Video-LLMs. *arXiv preprint arXiv:2406.07476* (2024). <https://arxiv.org/abs/2406.07476>
- [13] Quang Minh Dinh, Minh Khoi Ho, Anh Quan Dang, and Hung Phong Tran. 2024. TrafficVLM: A Controllable Visual Language Model for Traffic Video Captioning. *arXiv:2404.09275 [cs.CV]* <https://arxiv.org/abs/2404.09275>
- [14] Shiloh L Dockstader and A Murat Tekalp. 2001. Multiple camera tracking of interacting and occluded human motion. *Proc. IEEE* 89, 10 (2001), 1441–1455. doi:10.1109/5.959341
- [15] Parya Dolatyabi, Jacob Regan, and Mahdi Khodayar. 2025. Deep Learning for Traffic Scene Understanding: A Review. *IEEE Access* 13 (2025), 13187–13237. doi:10.1109/ACCESS.2025.3529289
- [16] Kai Du, Xingping Guo, Letian Li, Jingni Song, Qingqing Shi, Mengyao Hu, and Jianwu Fang. 2025. Traffic prediction in time series, spatialtemporal, and OD data: A systematic survey. *Journal of Traffic and Transportation Engineering (English Edition)* (2025). doi:10.1016/j.jtte.2025.03.001
- [17] Wenlu Du, Ankan Dash, Jing Li, Hua Wei, and Guiling Wang. 2023. Safety in Traffic Management Systems: A Comprehensive Survey. *Designs* 7 (08 2023), 100. doi:10.3390/designs7040100
- [18] Leonard Evans. 1991. *Traffic Safety and the Driver*. Van Nostrand Reinhold, New York, NY.
- [19] Ziang Guo, Zakhar Yagudin, Artem Lykov, Mikhail Konenkov, and Dzmitry Tsetserukou. 2024. VLM-Auto: VLM-based Autonomous Driving Assistant with Human-like Behavior and Understanding for Complex Road Scenes. *arXiv*

- preprint arXiv:2405.05885* (2024). <https://arxiv.org/abs/2405.05885>
- [20] Ismail Haritaoglu, David Harwood, and Larry S Davis. 2000. W4: Real-time surveillance of people and their activities. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 22, 8 (2000), 809–830. doi:10.1109/34.868683
- [21] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, and Weizhu Chen. 2021. LoRA: Low-Rank Adaptation of Large Language Models. *arXiv preprint arXiv:2106.09685* (2021).
- [22] Jyh-Jing Hwang, Runsheng Xu, Hubert Lin, Wei-Chih Hung, Jingwei Ji, Kristy Choi, Di Huang, Tong He, Paul Covington, Benjamin Sapp, Yin Zhou, James Guo, Dragomir Anguelov, and Mingxing Tan. 2024. EMMA: End-to-End Multimodal Model for Autonomous Driving. *arXiv preprint arXiv:2410.23262* (2024). <https://arxiv.org/abs/2410.23262>
- [23] Muhammad Monjurul Karim, Yan Shi, Shucheng Zhang, Bingzhang Wang, Mehrdad Nasri, and Yinhai Wang. 2025. Large Language Models and Their Applications in Roadway Safety and Mobility Enhancement: A Comprehensive Review. *arXiv:2506.06301* [cs.AI] <https://arxiv.org/abs/2506.06301>
- [24] Quan Kong, Yuki Kawana, Rajat Saini, Ashutosh Kumar, Jingjing Pan, Ta Gu, Yohei Ozao, Balazs Opra, Yoichi Sato, and Norimasa Kobori. 2024. WTS: A Pedestrian-Centric Traffic Video Dataset for Fine-grained Spatial-Temporal Understanding. In *European Conference on Computer Vision (ECCV)*. [https://www.ecva.net/papers/eccv\\_2024/papers\\_ECCV/papers/09667.pdf](https://www.ecva.net/papers/eccv_2024/papers_ECCV/papers/09667.pdf)
- [25] Quan KongKong et al., Yuki Kawana, Rajat Saini, Ashutosh Kumar, Jingjing Pan, Ta Gu, Yohei Ozao, Balazs Opra, and David C Anastasiu. 2024. WTS: A Pedestrian-Centric Traffic Video Dataset for Fine-grained Spatial-Temporal Understanding. *arXiv preprint arXiv:2407.15350* (2024). arXiv:2407.15350 [cs.CV]
- [26] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. 2017. Dense-captioning events in videos. In *Proceedings of the IEEE International Conference on Computer Vision*. IEEE, 706–715. doi:10.1109/ICCV.2017.83
- [27] I. Laia, J.J. Sanchez-Medina, E.I. Vlahogianni, and J. Del Ser. 2021. From Data to Actions in Intelligent Transportation Systems: A Prescription of Functional Requirements for Model Actionability. *Sensors* 21, 4 (2021), 1121. doi:10.3390/s21041121
- [28] Boyi Li, Ligeng Zhu, Ran Tian, Shuhan Tan, Yuxiao Chen, Yao Lu, Yin Cui, Sushant Veer, Max Ehrlich, Jonah Philion, Xinshuo Weng, Fuzhao Xue, Andrew Tao, Ming-Yu Liu, Sanja Fidler, Boris Ivanovic, Trevor Darrell, Jitendra Malik, Song Han, and Marco Pavone. 2024. Wolf: Captioning Everything with a World Summarization Framework. *arXiv preprint arXiv:2407.18908* (2024). <https://arxiv.org/abs/2407.18908>
- [29] Zihao Li, Chaolun Ma, Yang Zhou, Dominique Lord, and Yunlong Zhang. 2025. Leveraging Textual Description and Structured Data for Estimating Crash Risks of Traffic Violation: A Multimodal Learning Approach. *IEEE Transactions on Intelligent Transportation Systems* 26, 7 (2025), 9201–9213. doi:10.1109/TITS.2025.3568287
- [30] Bin Lin, Bin Zhu, Yang Ye, Munan Ning, Peng Jin, and Li Yuan. 2023. Video-LLaVA: Learning United Visual Representation by Alignment Before Projection. *ArXiv abs/2311.10122* (2023). <https://api.semanticscholar.org/CorpusID:265281544>
- [31] Bin Lin, Bin Zhu, Yang Ye, Munan Ning, Peng Jin, and Li Yuan. 2023. Video-LLaVA: Learning United Visual Representation by Alignment Before Projection. *arXiv preprint arXiv:2311.10122* (2023).
- [32] Fugang Liu, Songnan Duan, and Wang Juan. 2024. A pedestrian trajectory prediction method based on improved LSTM network. *IET Image Processing* 18, 2 (2024), 379–387. doi:10.1049/ipt2.12954
- [33] Sheng Liu, Haotian Ye, Lei Xing, and James Zou. 2024. Reducing Hallucinations in Vision-Language Models via Latent Space Steering. *arXiv preprint arXiv:2410.15778* (2024). <https://arxiv.org/abs/2410.15778>
- [34] Fan Ma, Xiaojie Jin, Heng Wang, Yuchen Xian, Jiashi Feng, and Yi Yang. 2024. VISTA-LLAMA: Reducing Hallucination in Video Language Models via Equal Distance to Visual Tokens. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 13151–13160. [https://openaccess.thecvf.com/content/CVPR2024/papers/Ma\\_VISTA-LLAMA\\_Reducing\\_Hallucination\\_in\\_Video-Language\\_Models\\_via\\_Equal\\_Distance\\_CVPR\\_2024\\_paper.pdf](https://openaccess.thecvf.com/content/CVPR2024/papers/Ma_VISTA-LLAMA_Reducing_Hallucination_in_Video-Language_Models_via_Equal_Distance_CVPR_2024_paper.pdf)
- [35] Arash Mahboubi, Khanh Luong, Hamed Abutorab, Hang Thanh Bui, Geoff Jarrad, Mohammed Bahatur, Seyit Camtepe, Ganna Pogrebna, Ejaz Ahmed, Bazara Barry, and Hannah Gately. 2024. Evolving techniques in cyber threat hunting: A systematic review. *Journal of Network and Computer Applications* 232 (2024), 104004. doi:10.1016/j.jnca.2024.104004
- [36] Ana-Maria Marcu, Long Chen, Jan Hünermann, Alice Karnsund, Benoit Hanotte, Prajwal Chidananda, Saurabh Nair, Vijay Badrinarayanan, Alex Kendall, Jamie Shotton, and Oleg Sinavskii. 2023. LingoQA: Visual Question Answering for Autonomous Driving. *arXiv preprint arXiv:2312.14115* (2023).
- [37] Alireza Mirzabagheri, Majid Ahmadi, Ning Zhang, Reza Alirezaei, Saeed Mozafari, and Shahpour Alirezaei. 2025. Navigating Uncertainty: Advanced Techniques in Pedestrian Intention Prediction for Autonomous Vehicles—A Comprehensive Review. *Vehicles* 7 (06 2025), 57. doi:10.3390/vehicles7020057
- [38] Tanner Muturi, Blessing Agyei Kyem, Joshua Kofi Asamoah, Jakisa Neema Owor, Richard Dyzinela, Andrews Danyo, Yaw Adu-Gyamfi, and Armstrong Aboah. 2025. Prompt-Guided Spatial Understanding with RGB-D Transformers for Fine-Grained Object Relation Reasoning. In *The IEEE International Conference on Computer Vision (ICCV) Workshops*.
- [39] Alexandru Niculescu-Mizil, Deep Patel, and Iain Melvin. 2024. MCTR: Multi Camera Tracking Transformer. *arXiv:2408.13243* [cs.CV] <https://arxiv.org/abs/2408.13243>
- [40] Murat Arda Onsu, Poonam Lohan, Burak Kantarci, Aisha Syed, Matthew Andrews, and Sean Kennedy. 2025. Leveraging Multimodal-LLMs Assisted by Instance Segmentation for Intelligent Traffic Monitoring. *arXiv:2502.11304* [cs.AI] <https://arxiv.org/abs/2502.11304>
- [41] Neema Jakisa Owor, Joshua Kofi Asamoah, Tanner Muturi, Jakisa Anneliese Owor, Blessing Agyei Kyem, Andrews Danyo, Yaw Adu-Gyamfi, and Armstrong Aboah. 2025. A Unified Detection Pipeline for Robust Object Detection in Fisheye-Based Traffic Surveillance. In *The IEEE International Conference on Computer Vision (ICCV) Workshops*.
- [42] Gracile Astlin Pereira and Muhammad Hussain. 2024. A Review of Transformer-Based Models for Computer Vision Tasks: Capturing Global Context and Spatial Relationships. *arXiv:2408.15178* [cs.CV] <https://arxiv.org/abs/2408.15178>
- [43] Tianwen Qian, Jingjing Chen, Linhai Zhuo, Yang Jiao, and Yu-Gang Jiang. 2023. NuScenes-QA: A Multi-modal Visual Question Answering Benchmark for Autonomous Driving Scenario. *arXiv preprint arXiv:2305.14836* (2023).
- [44] Alex Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandini Agarwal, Gabriel Brockman, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. *arXiv preprint arXiv:2103.00020* (2021).
- [45] Adib Bin Rashid and MD Ashfakul Karim Kausik. 2024. AI revolutionizing industries worldwide: A comprehensive overview of its diverse applications. *Hybrid Advances* 7 (2024), 100277. doi:10.1016/j.hybadv.2024.100277
- [46] James Reason. 1997. *Managing the Risks of Organizational Accidents*. Ashgate Publishing, Aldershot, UK.
- [47] Angela Schubert, Stefan Babisch, John M. Scanlon, Eamon T. Campolettano, Robby Roessler, Thomas Unger, and Timothy L. McMurry. 2023. Passenger and heavy vehicle collisions with pedestrians: Assessment of injury mechanisms and risk. *Accident Analysis & Prevention* 190 (2023), 107139. doi:10.1016/j.aap.2023.107139
- [48] Liang Shi, Boyu Jiang, Tong Zeng, and Feng Guo. 2024. ScVLM: Enhancing Vision-Language Model for Safety-Critical Event Understanding. *arXiv preprint arXiv:2410.00982* (2024). <https://arxiv.org/abs/2410.00982>
- [49] Maged Shoman, Dongdong Wang, Armstrong Aboah, and Mohamed Abdel-Aty. 2024. Enhancing Traffic Safety with Parallel Dense Video Captioning for End-to-End Event Analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. 7125–7133. [https://openaccess.thecvf.com/content/CVPR2024W/AICity/papers/Shoman\\_Enhancing\\_Traffic\\_Safety\\_with\\_Parallel\\_Dense\\_Video\\_Captioning\\_for\\_End-to-End\\_CVPRW\\_2024\\_paper.pdf](https://openaccess.thecvf.com/content/CVPR2024W/AICity/papers/Shoman_Enhancing_Traffic_Safety_with_Parallel_Dense_Video_Captioning_for_End-to-End_CVPRW_2024_paper.pdf)
- [50] Maged Shoman, Dongdong Wang, Armstrong Aboah, and Mohamed Abdel-Aty. 2024. Enhancing Traffic Safety with Parallel Dense Video Captioning for End-to-End Event Analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*. IEEE, 3521–3529. doi:10.1109/CVPRW.2024.000678
- [51] Chonghao Sima, Katrin Renz, Kashyap Chitta, Li Chen, Hanxue Zhang, Chengen Xie, Ping Luo, Andreas Geiger, and Hongyang Li. 2024. DriveLM: Driving with Graph Visual Question Answering. In *European Conference on Computer Vision (ECCV)*.
- [52] Z. Sun et al. 2025. Multiple Pedestrian Tracking Under Occlusion: A Survey and Outlook. *IEEE Transactions on Circuits and Systems for Video Technology* 35, 2 (2025), 1009–1027. doi:10.1109/TCST.2024.3481425
- [53] Qwen Team. 2024. Qwen2 Technical Report. *arXiv:2407.10671* [cs.CL] <https://arxiv.org/abs/2407.10671>
- [54] Xiaoyu Tian, Junru Gu, Bailin Li, Yicheng Liu, Yang Wang, Zhiyong Zhao, Kun Zhan, Peng Jia, Xianpeng Lang, and Hang Zhao. 2024. DriveVLM: The Convergence of Autonomous Driving and Large Vision-Language Models. *arXiv preprint arXiv:2402.12289* (2024). <https://arxiv.org/abs/2402.12289>
- [55] Tuan-An To, Minh-Nam Tran, Trong-Bao Ho, Thien-Loc Ha, Quang-Tan Nguyen, Hoang-Chau Luong, Thanh-Duy Cao, and Minh-Triet Tran. 2024. Multi-perspective Traffic Video Description Model with Fine-grained Refinement Approach. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. 7075–7084. [https://openaccess.thecvf.com/content/CVPR2024W/AICity/papers/To\\_Multi-perspective\\_Traffic\\_Video\\_Description\\_Model\\_with\\_Fine-grained\\_Refinement\\_Approach\\_CVPRW\\_2024\\_paper.pdf](https://openaccess.thecvf.com/content/CVPR2024W/AICity/papers/To_Multi-perspective_Traffic_Video_Description_Model_with_Fine-grained_Refinement_Approach_CVPRW_2024_paper.pdf)
- [56] Shuo Wang, David C. Anastasiu, Zheng Tang, Ming-Ching Chang, Yue Yao, Liang Zheng, Mohammed Shaiqur Rahman, Meenakshi S. Arya, Anuj Sharma, Pranamesh Chakraborty, Sanjana Prajapati, Quan Kong, Norimasa Kobori, Munkh-jargal Gochoo, Munkh-Erdene Otgonbold, Fady Alnajjar, Ganzorig Batnasan, Ping-Yang Chen, Jun-Wei Hsieh, Xunlei Wu, Sameer Satish Pusegaonkar, Yizhou Wang, Sujit Biswas, Rama Chellappa, and .... 2024. The 8th AI City Challenge. <https://arxiv.org/abs/2404.09432> arXiv preprint arXiv:2404.09432.

- [57] Yueyang Wang, Aravinda Ramakrishnan Srinivasan, Jussi P.P. Jokinen, Antti Oulasvirta, and Gustav Markkula. 2025. Pedestrian crossing decisions can be explained by bounded optimal decision-making under noisy visual perception. *Transportation Research Part C: Emerging Technologies* 171 (2025), 104963. doi:10.1016/j.trc.2024.104963
- [58] Siyuan Wu, Xiaobing Zhang, Xin Wang, Chen Li, and Licheng Jiao. 2020. Scene attention mechanism for remote sensing image caption generation. In *2020 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 1–7. doi:10.1109/IJCNN48605.2020.9207326
- [59] Chen-Wei Xie, Siyang Sun, Xiong Xiong, Yun Zheng, Deli Zhao, and Jingren Zhou. 2023. Retrieval-Augmented Contrastive Language-Image Pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- [60] Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, et al. 2024. MiniCPM-V: A GPT-4V Level MLLM on Your Phone. *arXiv preprint arXiv:2408.01800* (2024).
- [61] Boqiang Zhang, Kehan Li, Zesen Cheng, Zhiqiang Hu, Yuqian Yuan, Guanzheng Chen, Sicong Leng, Yuming Jiang, Hang Zhang, Xin Li, Peng Jin, Wenqi Zhang, Fan Wang, Lidong Bing, and Deli Zhao. 2025. VideoLLaMA 3: Frontier Multi-modal Foundation Models for Image and Video Understanding. *arXiv preprint arXiv:2501.13106* (2025). <https://arxiv.org/abs/2501.13106>
- [62] Hang Zhang, Xin Li, and Lidong Bing. 2023. Video-LLaMA: An Instruction-tuned Audio-Visual Language Model for Video Understanding. 543–553. doi:10.18653/v1/2023.emnlp-demo.49
- [63] Hang Zhang, Xin Li, and Lidong Bing. 2023. Video-LLaMA: An Instruction-tuned Audio-Visual Language Model for Video Understanding. *arXiv preprint arXiv:2306.02858* (2023). <https://arxiv.org/abs/2306.02858>
- [64] Ruixuan Zhang, Beichen Wang, Juxiao Zhang, Zilin Bian, Chen Feng, and Kaan Ozbay. 2025. When language and vision meet road safety: leveraging multimodal large language models for video-based traffic accident analysis. *arXiv preprint arXiv:2501.10604* (2025).
- [65] Yuanhan Zhang, Bo Li, haotian Liu, Yong jae Lee, Liangke Gui, Di Fu, Jiashi Feng, Ziwei Liu, and Chunyuan Li. 2024. LLaVA-NeXT: A Strong Zero-shot Video Understanding Model. <https://llava-vl.github.io/blog/2024-04-30-llava-next-video/>
- [66] Zhaoyang Zhang, Yantao Shen, Kunyu Shi, Zhaowei Cai, Jun Fang, Siqi Deng, Hao Yang, Davide Modolo, Zhiuwen Tu, and Stefano Soatto. 2023. Musketeer: Joint Training for Multi-Task Vision-Language Model with Task Explanation Prompts. *arXiv preprint arXiv:2305.07019* (2023).
- [67] Rui Zhao, Qirui Yuan, Jinyu Li, Yuze Fan, Yun Li, and Fei Gao. 2024. DriveLLaVA: Human-Level Behavior Decisions via Vision Language Model. *Sensors* 24, 13 (2024), 4113. doi:10.3390/s24134113
- [68] Luowei Zhou, Yingbo Zhou, Jason J Corso, Richard Socher, and Caiming Xiong. 2018. End-to-end dense video captioning with masked transformer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 8739–8748. doi:10.1109/CVPR.2018.00911
- [69] Wei Zhou, Lei Zhao, Runyu Zhang, Yifan Cui, Hongpu Huang, Kun Qie, and Chen Wang. 2024. Vision Technologies with Applications in Traffic Surveillance Systems: A Holistic Survey. *arXiv:2412.00348 [cs.CV]* <https://arxiv.org/abs/2412.00348>

Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009