Arab American University
Faculty of Graduate Studies
Department of Natural, Engineering and Technology Sciences
Master Program in Data Science and Business Analytics

Enhancing Traffic Safety with a Multimodal Large Language Model for Real-Time Hazard Detection

Mohammad Yaser Ahmad Abu Tami
202216440

Supervision Committee:
Dr. Mohammed Elhenawy
Dr. Huthaifa I. Ashqar
Dr. Ahmad Hasasneh

This Thesis Was Submitted in Partial Fulfillment of the Requirements for the Master Degree in
Data Science and Business Analytics

Palestine, July/ 2025

Arab American University
Faculty of Graduate Studies
Department of Natural, Engineering and Technology Sciences
Master Program in Data Science and Business Analytics

Thesis Approval

Enhancing Traffic Safety with a Multimodal Large Language Model for Real-Time Hazard Detection

Mohammad Yaser Ahmad Abu Tami
202216440

This thesis was defended successfully on 27th July 2025 and approved by:

Thesis Committee Members:

	Name	Title	Signature
1.	Dr. Mohammed Elhenawy	Main Supervisor	
2.	Dr. Huthaifa I. Ashqar	Members of Supervision Committee	
3.	Dr. Ahmad Hasasneh	Members of Supervision Committee	

Palestine, July/ 2025

Declaration

I declare that, except where explicit reference is made to the contribution of others, this thesis is substantially my own work and has not been submitted for any other degree at the Arab American University or any other institution.

Student Name: Mohammad Yaser Ahmad Abu Tami

Student ID: 202216440

Signature:

Date of Submitting the Final Version of the Thesis: 14.8.2025

Dedication

To my wife, Abrar, whose love, patience, and unwavering belief in me have been my greatest source of strength.

To my family, whose unwavering support and encouragement have been my guiding light throughout this journey.

To my mentors and teachers, who inspired me to pursue knowledge and strive for excellence.

To my people, the Palestinians, whose resilience, courage, and unwavering spirit continue to inspire hope and determination.

And to all those who believe in the power of perseverance and the pursuit of dreams, this work is dedicated to you.

Mohammad Yaser Ahmad Abu Tami

Acknowledgments

I would like to express my deepest gratitude to my supervisors, examiners, fellow students, technical staff, and everyone who contributed to the completion of this thesis. Their invaluable guidance, constructive feedback, and generous support have greatly enriched the quality of my work.

I am especially thankful to the Arab American University, acknowledging the vital role of the Faculty, Department, program coordinator, and program committee. Their dedication, the availability of research facilities, and the supportive academic environment provided me with the tools and encouragement needed to carry out this research to the best of my ability.

I extend my sincere appreciation to the Honda Research Laboratory for providing access to the dataset used in this work. This dataset corresponds to the paper titled *"DRAMA: Joint Risk Localization and Captioning in Driving"* by Malla et al., presented at the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) 2023. The availability of this resource has been instrumental in advancing my research, enabling the development and validation of key methodologies presented in this thesis.

Finally, I am grateful to my family and friends for their unwavering moral support, patience, and encouragement throughout this journey. Their belief in me has been a source of motivation from start to finish.

Enhancing Traffic Safety with a Multimodal Large Language Model for Real-Time Hazard Detection

Mohammad Yaser Ahmad Abu Tami

Dr. Mohammed Elhenawy
Dr. Huthaifa I. Ashqar
Dr. Ahmad Hasasneh

Abstract

Traffic safety remains a critical global issue, with traditional detection systems often falling short in complex, real-world environments due to limited generalizability and high computational demands. This thesis introduces HazardNet, a lightweight, edge-compatible Multimodal Large Language Model (MLLM) fine-tuned from Qwen2-VL-2B using parameter-efficient techniques like Low-Rank Adaptation (LoRA) and Quantized Low-Rank Adaptation (QLoRA). HazardNet is designed for real-time, contextual hazard detection by combining visual and textual inputs, making it suitable for deployment on low-resource devices such as in-vehicle systems without GPUs.

To support this model, the study presents HazardQA, a new Vision Question Answering (VQA) dataset derived from driving scenarios in the DRAMA dataset. HazardQA includes over 7,000 annotated question-answer pairs enriched with reasoning chains and safety-specific labels, covering a wide range of hazards and traffic contexts.

Experiments show that HazardNet performs competitively on safety-critical tasks such as scene understanding, hazard identification, and action recommendation, rivaling larger models like GPT-4o while maintaining low computational requirements. The research highlights how compact MLLMs, when adapted with domain-specific data and efficient fine-tuning methods, can provide high interpretability and scalability for traffic safety systems. Key contributions include the development of HazardNet, the open-source release of HazardQA, and a demonstration of real-world deployment strategies for AI-based hazard detection.

Keywords: Traffic Safety, Multimodal Language Models, Vision Question Answering, Low-Rank Adaptation, Edge Deployment.

Table of Contents

List of Tables

List of Figures

List of Appendices

List of Definitions of Abbreviations

Abbreviations	Title
AD	Autonomous Driving
ADAS	Advanced Driver-Assistance Systems
AI	Artificial Intelligence
AV	Automated Vehicle
BERT	Bidirectional Encoder Representations from Transformer
CNN	Convolutional Neural Network
CoT	Chain-of-Thought
GNN	Graph Neural Network
GPT	Generative Pre-Training Transformer
ICL	In-Context Learning
LLaVA	Large Language and Vision Assistant
LLM	Large Language Model
LoRA	Low-Rank Adaptation
M-ROPE	Multimodal Rotary Position Embedding
ML	Machine Learning
MLLM	Multimodal Large Language Model
NLP	Natural Language Processing
QLoRA	Quantized Low-Rank Adaptation
SCE	Safety-Critical Event
SOTA	State of The Art
VFM	Vision Foundation Model
ViT	Vision Transformers
VLM	Vision Language Model

VQA	Vision Question Answering
WHO	World Health Organization

Chapter One: Introduction

1.1 Traffic Safety and Machine Learning

Traffic safety remains a critical global concern, with millions of accidents occurring annually, leading to significant loss of life and economic costs. According to the World Health Organization (WHO), road traffic injuries are the leading cause of death for children and young adults aged 5–29 years (World Health Organizatio, 2023). This highlighting the urgent need for innovative solutions to enhance road safety. Traditional approaches to traffic safety have relied on rule-based systems and classical machine learning (ML) models, such as decision trees, support vector machines, and convolutional neural networks (CNNs) (Komasi et al., 2024; Sohail et al., 2023). While these methods have achieved some success, they face inherent limitations in addressing the dynamic, unpredictable nature of real-world driving scenarios. For example, they often require extensive annotated datasets, struggle to generalize across diverse environments (e.g., urban intersections, highways, rural roads), and lack the ability to reason about complex, context-dependent events (Bansal et al., 2020).

Artificial Intelligence (AI) and Machine Learning (ML) have emerged as transformative technologies with the potential to address complex challenges in traffic safety (Pérez-Castán et al., 2022). AI refers to the simulation of human intelligence in machines, enabling them to perform tasks such as perception, reasoning, and decision-making (Jarrahi, 2018). Machine Learning, a subset of AI, involves training algorithms to learn patterns from data and make predictions or decisions without explicit programming (Mahesh, 2020). Over the past decade, advancements in AI and ML have led to significant improvements in areas such as computer vision, natural language processing (NLP), and multimodal learning, which combine multiple data modalities (e.g., text, images, and sensor data) to solve complex problems (Nam & Jang, 2024).

In the context of traffic safety, AI and ML have been applied to tasks such as object detection, anomaly detection, and driver behavior analysis. For example, computer vision models can identify pedestrians, vehicles, and traffic signs in real-time, while anomaly detection algorithms can flag unsafe driving behaviors or hazardous road conditions (Khan et al., 2023). However, despite these advancements, existing solutions face several challenges, including the need for large annotated datasets, limited generalization across diverse environments, and high computational requirements, which hinder their deployment in resource-constrained settings (Abu Tami et al., 2024).

One of the key challenges in developing effective traffic safety systems is the ability to detect and respond to safety-critical events in real-time. Safety-critical events, such as sudden braking, pedestrian crossings, or vehicle collisions, require immediate attention and often occur in complex, multi-modal environments that involve both visual and contextual information. While recent advancements in AI, such as deep learning and transformer-based models, have shown promise in addressing these challenges, their high computational requirements and reliance on powerful hardware limit their applicability in real-world, resource-constrained settings, such as edge devices or low-power systems (Abibullaev et al., 2023; S. Liu et al., 2019).

To address these challenges, this research proposes HazardNet, a small-scale Multimodal Language Model (MLM) designed to enhance traffic safety by leveraging the reasoning capabilities of advanced language and vision-language models. HazardNet aims to detect safety-critical events in real-time while maintaining computational efficiency, making it suitable for deployment in resource-constrained environments. By combining the strengths of multimodal learning and compact model architectures, HazardNet seeks to overcome the limitations of existing solutions and provide a practical, scalable approach to improving traffic safety.

The following sections will discuss the problem overview, research objectives, and methodology, providing a comprehensive foundation for this study. By leveraging the latest advancements in MLLMs, this research seeks to contribute to the development of intelligent systems that can anticipate, detect, and mitigate hazards, ultimately reducing the global burden of traffic-related injuries and fatalities.

1.2 Problem Overview

Traffic safety is a critical issue in rapidly urbanizing areas. Effective traffic safety management not only saves lives but also reduces economic losses associated with accidents, such as healthcare costs, property damage, and loss of productivity. Traditional safety-critical event detection systems predominantly rely on sensor-based approaches and conventional ML algorithms, which demand extensive annotated data and struggle to generalize across the diverse and dynamic conditions of urban traffic. These systems often operate in isolation, limiting their effectiveness in comprehensive traffic management.

Recent advancements in Large Language Models (LLMs) and Vision Language Models (VLMs) have introduced promising enhancements for traffic safety systems by integrating multiple data modalities, such as text and images, to improve event detection and decision-making. However, the large size and computational demands of these models hinder their deployment on edge devices, which are necessary for real-time applications (Abu Tami et al., 2024). This gap highlights the need for compact, efficient models that can operate effectively in resource-constrained environments while maintaining high accuracy and reliability.

The current landscape of traffic safety systems reveals a critical gap: the lack of compact, efficient, and scalable models that can leverage the multimodal capabilities of LLMs and VLMs while operating effectively in resource-constrained environments. Existing solutions either rely on traditional ML approaches with limited generalization capabilities or employ advanced LLMs/VLMs that are impractical for real-time, edge-based deployment. Bridging this gap requires innovative approaches that balance model complexity, computational efficiency, and accuracy to enable the widespread adoption of intelligent traffic safety systems in urban environments.

1.3 Research Objectives

The primary objective of this research is to develop HazardNet, a small-scale Multimodal Language Model (MLM) designed to enhance traffic safety by leveraging the reasoning capabilities of advanced language and vision-language models. Multimodal Language Models, such as Qwen2-VL-2B and GPT-4o, combine text and visual data to perform tasks like Visual Question Answering (VQA), making them ideal for analyzing complex real-world scenarios. HazardNet focuses on detecting safety-critical events in driving environments, aiming to provide efficient and real-time solutions for resource-constrained settings, such as edge devices. Specifically, the study aims to:

1. Fine-tune the pre-trained Qwen2-VL-2B model (P. Wang et al., 2024): The Qwen2-VL-2B model, a state-of-the-art open-source vision-language model with two

billion parameters, is selected for its superior performance among open-source alternatives and its compact size. This research fine-tunes the model to enhance its ability to detect safety-critical events, ensuring efficient inference throughput for deployment on edge devices.

2. Construct HazardQA, a novel Vision Question Answering (VQA) dataset: A specialized dataset, HazardQA, is developed to train and evaluate HazardNet. This dataset includes annotated real-world driving scenarios involving safety-critical events, ensuring robust model training and generalization across diverse conditions.
3. Evaluate the performance of HazardNet in detecting safety-critical events: The performance of HazardNet is rigorously evaluated and compared to both the base model and larger models, such as GPT-4o, to demonstrate its effectiveness and efficiency. Metrics such as accuracy, inference speed, and resource utilization are used for this comparison.
4. Explore the feasibility of deploying HazardNet in low-resource environments: The study investigates the deployment of HazardNet on edge devices and evaluates the model's load performance in terms of throughput (tokens/second). This testing aims to assess the model's efficiency and adaptability to resource-constrained settings, ensuring its practicality for real-world applications.

1.4 Research Questions

To guide the research, the following questions are proposed:

1. How effective is HazardNet, a fine-tuned small-scale Multimodal Language Model, in detecting safety-critical events compared to its base model and larger models like GPT-4o?
2. What methodologies can be employed to construct and utilize HazardQA, a novel VQA dataset, for training HazardNet on real-world safety-critical scenarios?
3. What are the key challenges and limitations of deploying HazardNet on edge devices, and how can they be addressed to ensure efficient real-time performance?
4. How does HazardNet contribute to improving traffic safety management by providing accurate and timely detection of safety-critical events?

1.5 Research Methodology

The research methodology is illustrated in Figure 1.1 which structured into the following phases:

1. Model Selection and Fine-Tuning:

   The pre-trained Qwen2-VL-2B model is selected for its superior performance and compact size, making it suitable for resource-constrained environments. HazardNet is developed by fine-tuning this model to enhance its capabilities in detecting safety-critical events. This phase focuses on adapting the model to the specific requirements of traffic safety applications.

2. Dataset Construction:

   A novel Visual Question Answering (VQA) dataset, HazardQA, is constructed to train and evaluate HazardNet. This dataset includes diverse driving scenarios annotated with safety-critical events, ensuring robust model training. The dataset is designed to reflect real-world conditions, such as variations in lighting, weather, and traffic, to improve the model's generalization capabilities.

3. Experimental Evaluation:

   HazardNet is evaluated through extensive experiments to assess its performance in detecting safety-critical events. Key metrics such as accuracy, inference speed, and resource utilization are used to compare its performance with the base Qwen2-VL-2B model and larger models like GPT-4o. This phase validates the effectiveness of HazardNet in achieving high performance while maintaining computational efficiency.

4. Deployment and Performance Testing on Low-Resource Devices:

   HazardNet is deployed on low-resource devices, such as GPU-free systems or PCs with less powerful GPUs, to evaluate its real-time performance. The testing focuses on key metrics, including throughput (tokens per second), to assess the model's efficiency and suitability for resource-constrained environments. This phase ensures that HazardNet can operate effectively in real-world settings, making it a practical solution for enhancing traffic safety.

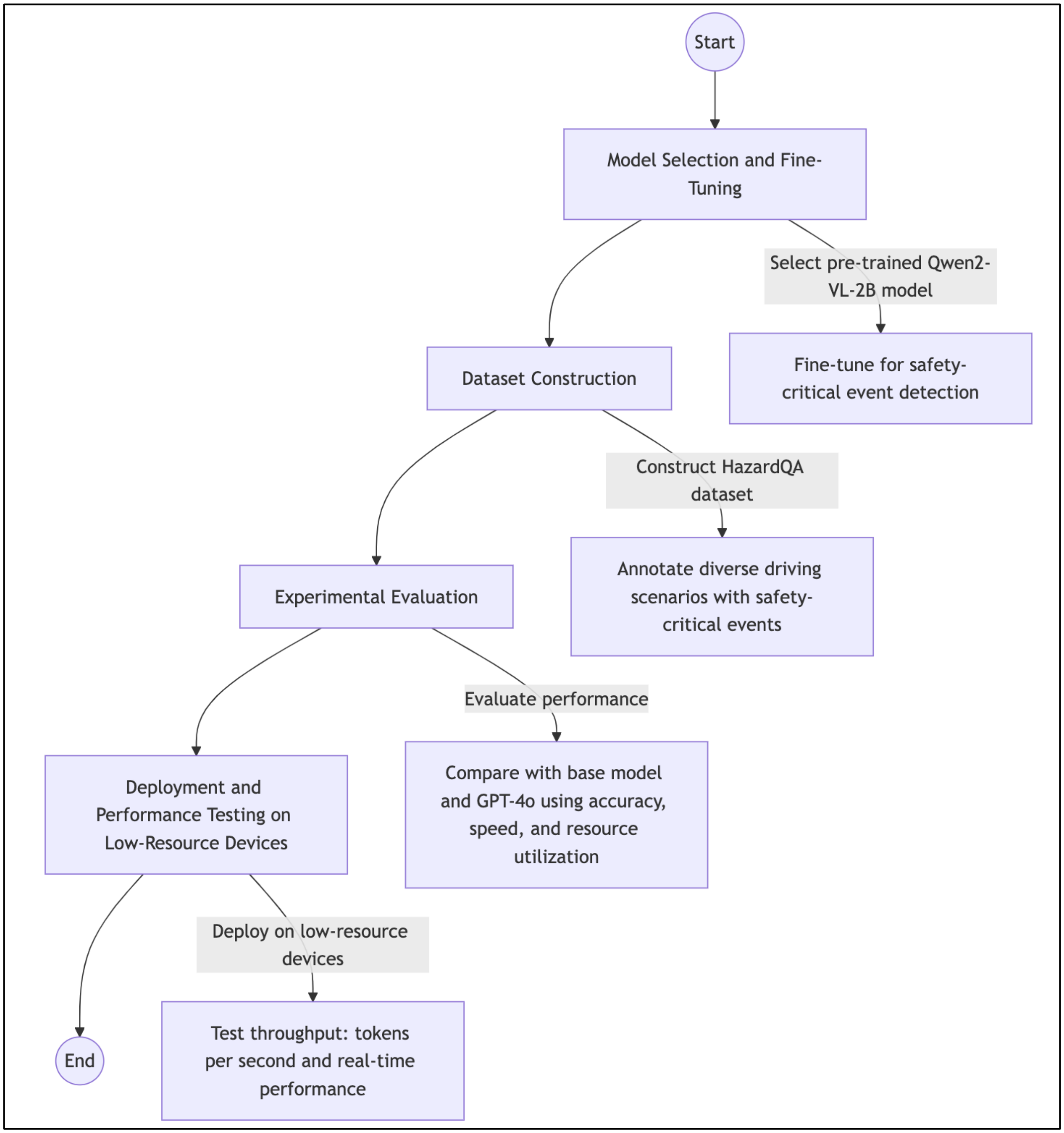


Figure 1.1 Workflow of HazardNet Development and Evaluation.

1.6 Thesis Organization

This thesis is organized into five chapters:

1. Introduction: Provides an overview of the research, including the problem statement, objectives, and methodology.

2. Background and Related Works: Examines existing research on traffic safety, multimodal language models, and edge computing, highlighting gaps and opportunities.
3. Methodology: Details the development and implementation of HazardNet, including model fine-tuning, dataset construction, and experimental design.
4. Results and Discussion: Presents the findings from the experiments, including performance metrics and comparative analysis, discusses their implications, and identifies limitations.
5. Conclusion and Future Works: Summarizes the key findings, contributions, and potential applications of HazardNet in enhancing traffic safety. Future works could explore the integration of real-time data streams from connected vehicles and IoT devices to further improve HazardNet's predictive capabilities. Additionally, investigating the use of advanced machine learning techniques, such as reinforcement learning or federated learning, could enhance the system's adaptability and scalability.

Chapter Two: Literature Review

2.1 Introduction

Traffic safety remains a critical global challenge, with road accidents causing substantial loss of life and economic damage annually (World Health Organization, 2023). Traditional hazard detection systems in Advanced Driver-Assistance Systems (ADAS) rely on isolated sensor technologies, such as LiDAR, cameras, and radar, paired with rule-based algorithms or conventional deep learning models. While effective in controlled environments, these systems struggle with dynamic real-world scenarios, particularly under adversarial environmental conditions like shadows, rain, fog, or sensor noise (Jaradat et al., 2024). For example, physical adversarial attacks, such as strategically placed perturbations on road signs or environmental distortions, can mislead object detection models, causing catastrophic failures in Automated Vehicles (AV) decision-making (Eykholt, 2018) (Ren et al., 2021) as illustrated in Figure 2.1. These vulnerabilities stem from the lack of holistic contextual reasoning in traditional systems, which process modalities (e.g., visual, spatial) in isolation, leading to fragmented interpretations (R. Zhang et al., 2025a).

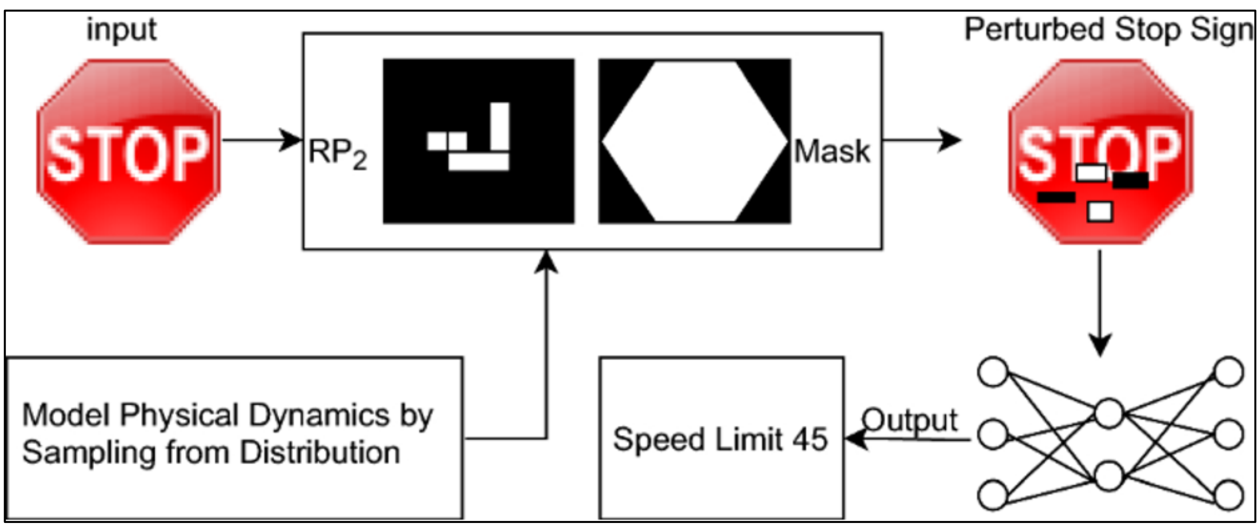


Figure 2.1 Illustration of an adversarial attack on a stop sign recognition system (Ren et al., 2021).

The emergence of Multimodal Large Language Models (MLLMs) offers transformative potential by integrating diverse data streams (visual, textual, auditory, environmental) into a unified reasoning framework. Unlike conventional models, MLLMs excel at correlating contextual cues, such as weather conditions, real-time traffic updates, and driver intent, to infer complex scenarios, including near-miss incidents or occluded pedestrian detection (S. Liu et al., 2024). Recent studies demonstrate that MLLMs enhance robustness against physical adversarial attacks by cross-validating sensor inputs. For instance, a distorted traffic sign detected by a camera can be contextualized using LiDAR distance data and weather sensors, mitigating misclassification risks (Y. Cao et al., 2019). Figure 2.2 contrasts traditional fragmented hazard detection with modern multimodal systems, emphasizing how integrated technologies like live tracking, context-aware radar, and environmental sensors enable cohesive scene understanding.

This work positions itself as a second layer atop existing AV frameworks, augmenting their safety through MLLM-driven contextual reasoning. By training on custom safety-critical events, such as sudden braking in foggy conditions or adversarial road artifacts, the proposed system addresses gaps in generalization and adversarial robustness. For non-automated vehicles, the framework enhances ADAS capabilities by interpreting driver behavior, road conditions, and external hazards in real time, bridging the gap between human intuition and machine precision (Kalamkar & A., 2023). For example, integrating MLLMs with ADAS can improve lane-keeping assistance during heavy rain by correlating windshield camera data with precipitation sensors and historical accident patterns (Zhou et al., 2024a). The integration of MLLMs into traffic safety systems offers several advantages:

1. Enhanced Contextual Understanding: MLLMs can process and correlate information from multiple modalities, such as video feeds, textual descriptions, and audio cues, to provide a holistic understanding of traffic scenarios (D. Zhang et al., 2024).
2. Real-Time Adaptability: By leveraging pre-trained models and fine-tuning techniques, MLLMs can be deployed on edge devices, enabling real-time hazard detection without the need for extensive computational resources (Mahmud et al., 2025).

3. Improved Generalization: MLLMs excel in zero-shot and few-shot learning, allowing them to generalize across diverse traffic conditions and scenarios, even with limited labeled data (Patil & Gudivada, 2024).

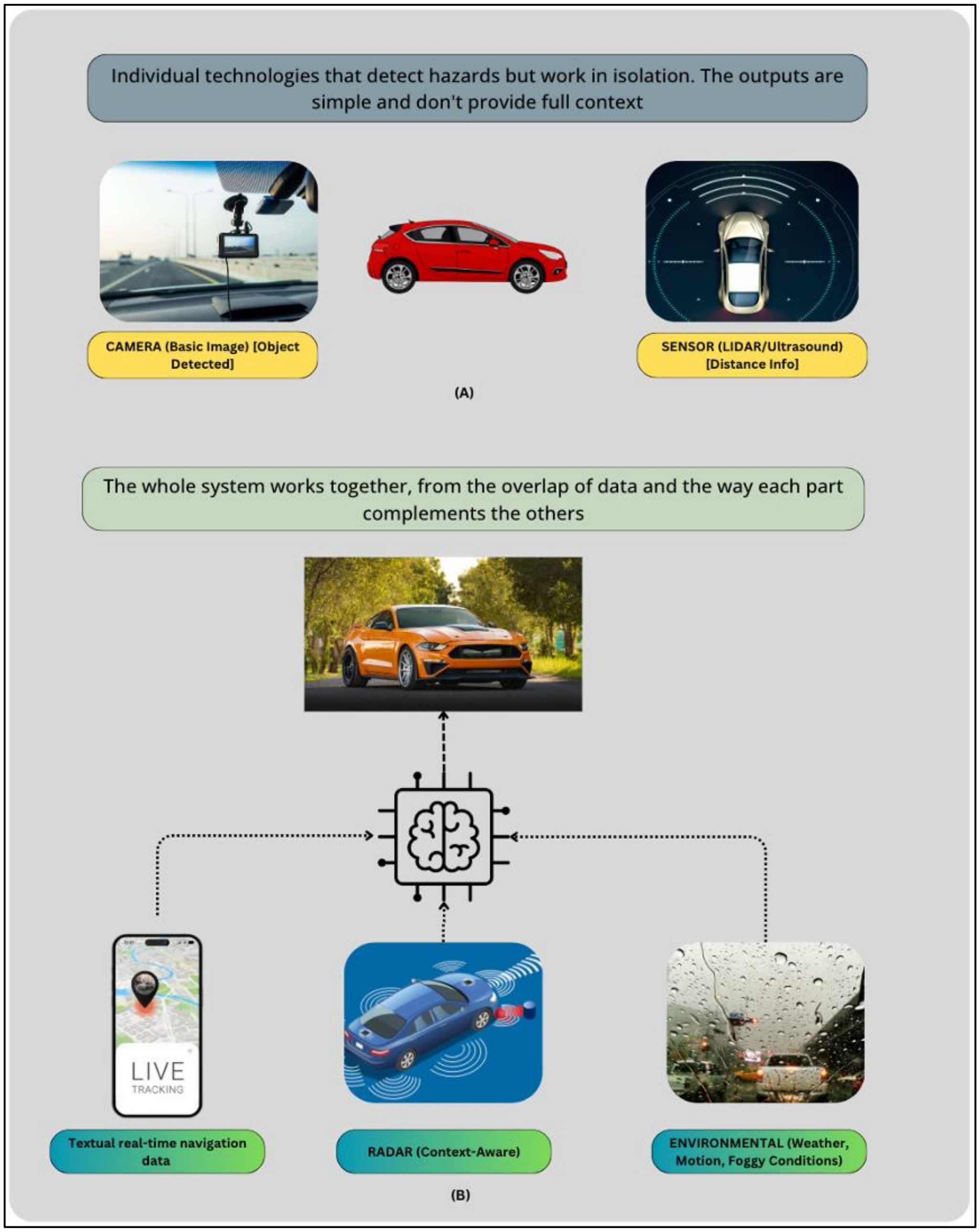


Figure 2.2 A conceptual illustration of traditional sensor-based hazard detection vs. multimodal AI-powered hazard detection systems.

2.2 Theoretical Background

2.2.1. Foundations of Transformer Architectures

A significant breakthrough in artificial intelligence (AI) came with the introduction of the Transformer architecture (Vaswani et al., 2017) as shown in Figure 2.3 which revolutionized natural language processing (NLP) by enabling models to process sequential data with unprecedented efficiency and accuracy. Transformers utilize solely the concept of self-attention mechanisms in addition to other state of the art (SOTA) techniques such as multi-head attention, scaled dot-product and positional encoding, allowing models to focus on relevant parts of the input sequence, thereby improving their ability to capture long-range dependencies and contextual relationships.

The main building block of the Transformer is the self-attention over the learnable parameters of the model, namely Query, Key and Values as illustrated in Figure 2.4 and expressed mathematically in equation 2.1. Self-attention computes the weighted sum of value (V) based on the similarity between query (Q) and key (K).

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^{\top}}{\sqrt{d_k}}\right)V \tag{2.1}$$

Where Q stands for Query, K for Key, and V for value. The dot product (QK^T) measures the similarity between query and key. Followed by normalizing the dot product using the key dimension to prevent the dominant of the big values. Finally, the SoftMax convert the scaled dot-product into probabilities.

The Introducing of the multi-head attention allow to parallelize the processing of the sequence data in the transformer layer, which help the model focus on different part of the sequence in each head. Mathematically the multi-heads can be expressed as shown in equation 2.2 and 2.3 respectively.

$$\begin{aligned} &MultiHead(Q, K, V) \\ &= \text{Concat}(\text{head}_1, \text{head}_2, \dots, \text{head}_h)W^O \end{aligned} \tag{2.2}$$

Where each head_i is computed as:

$$\text{head}_i = \text{Attention}\left(QW_i^Q, KW_i^K, VW_i^V\right) \tag{2.3}$$

Here:

- h is the number of attention heads.
- W_i^Q, W_i^K, W_i^V are learned projection matrices for the queries, keys, and values, respectively.
- W^O is the learned output projection matrix.

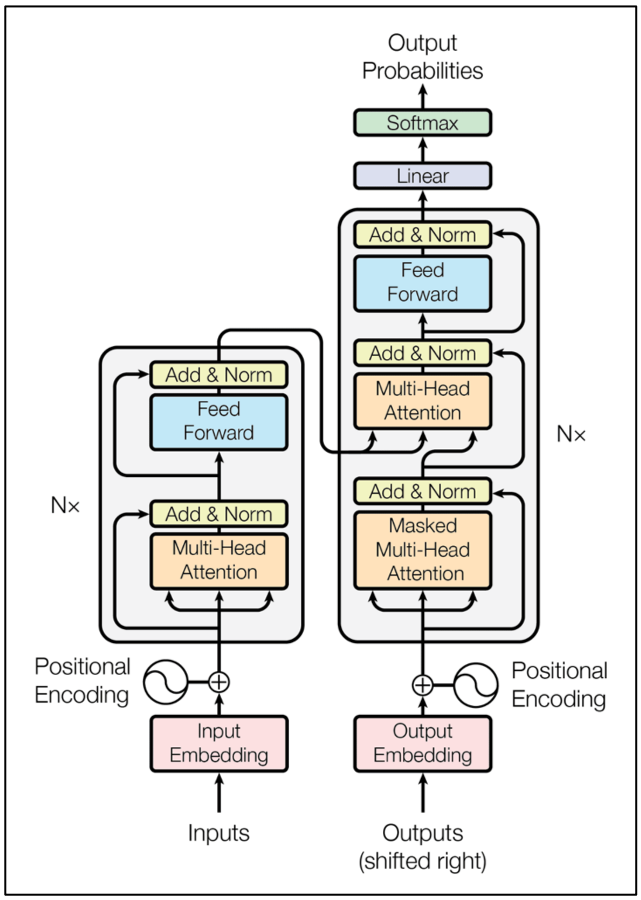


Figure 2.3 The Transformer architecture as proposed in the 'Attention is All You Need' paper. The model consists of an encoder-decoder structure, leveraging multi-head self-attention mechanisms and feedforward networks for sequence-to-sequence tasks (Vaswani et al., 2017).

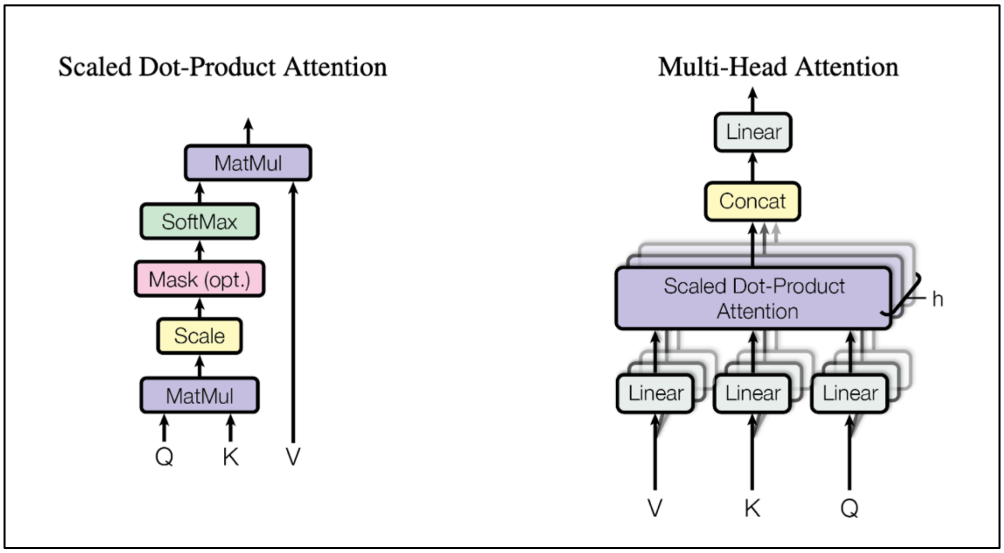


Figure 2.4 (Left) Scaled Dot-Product Attention: Computes attention weights using Q, K, V, with optional masking and scaling. (Right) Multi-Head Attention: Combines multiple attention heads for richer representation learning (Vaswani et al., 2017).

This innovation led to the development of large language models (LLMs) like Generative Pre-Training Transformer (GPT) (Brown et al., 2020)and Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2018), which have achieved state-of-the-art performance in a wide range of NLP tasks, including text generation, translation, and sentiment analysis.

Building on the success of Transformers in NLP, researchers extended this architecture to computer vision tasks, leading to the development of Vision Transformers (ViTs) (Dosovitskiy et al., 2020) . ViTs apply self-attention mechanisms to image patches, enabling them to process visual data with the same level of sophistication as text as shown in Figure 2.5 This breakthrough has significantly improved the performance of vision-based models in tasks such as object detection, image classification, and segmentation. Researchers have demonstrated that ViTs could outperform traditional convolutional neural networks (CNNs) on large-scale image recognition benchmarks, paving the way for their adoption in diverse applications, including autonomous driving and traffic safety (Khan et al., 2023), (Uparkar et al., 2023).

The integration of Transformers into multimodal learning frameworks has further expanded their potential. Multimodal Large Language Models (MLLMs) combine the strengths of NLP and computer vision, enabling the analysis of textual, visual, and auditory data within a unified architecture. Models like Gemini (Gemini Team et al., 2023) and Large Language and Vision Assistant (LLaVA) (H. Liu et al., 2023) have demonstrated remarkable capabilities in analyzing images. These models leverage advanced reasoning and contextual understanding to interpret multimodal inputs.

Recent studies have highlighted the potential of MLLMs in improving traffic safety. For example, Tami et al. (Abu Tami et al., 2024) proposed a framework that uses MLLMs to automate the detection of safety-critical events in driving videos, achieving promising results in zero-shot and few-shot learning scenarios. These advancements underscore the significance of MLLMs in advancing the analysis of naturalistic driving videos and improving the understanding of interactions within complex traffic environments.

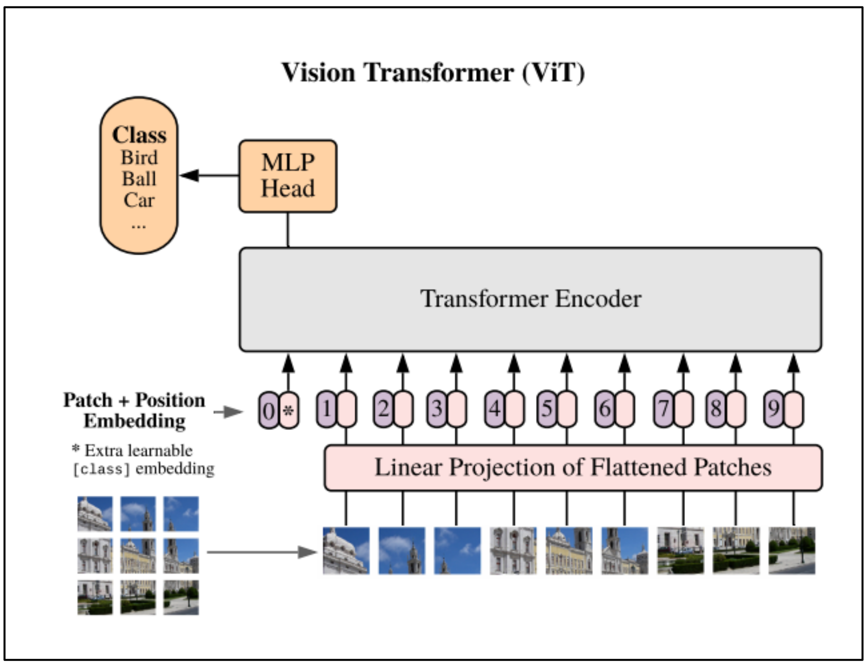


Figure 2.5 Vision Transformer (ViT) architecture with class tokens, MLP Head, Transformer Encoder, Patch + Position Embedding, and Linear Projection of Flattened Patches (Dosovitskiy et al., 2020).

2.2.2. Transfer Learning with Parameter Efficient Fine-Tuning Methods

While foundational models like Transformers, ViTs, and MLLMs achieve remarkable performance, their massive scale (often billions of parameters) poses challenges for continuous adaptation to new tasks or datasets. Full fine-tuning of all parameters is computationally prohibitive and risks *catastrophic forgetting*, where models lose previously learned knowledge. To address this, parameter-efficient tuning methods such as **Low-Rank Adaptation (LoRA)** (E. J. Hu et al., 2021) and **Quantized LoRA (QLoRA)** (Dettmers, 2024) have emerged, enabling efficient updates to pre-trained models while preserving their original capabilities.

Low-Rank Adaptation (LoRA) leverages the insight that weight updates (ΔW) during fine-tuning can be represented as low-rank matrices. For a pre-trained weight matrix $W \in R^{d\times k}$, LoRA constrains its update as:

$$\Delta W = B \cdot A \tag{2.4}$$

where $B \in R^{d\times r},\ A \in R^{r\times k},\ r \ll \min(d,k),$ effectively decomposing the update into two trainable low-rank matrices B and A. During training, W remains frozen, while only B and A are updated (Figure 2.6). This approach reduces the number of trainable parameters by orders of magnitude. For example, applying LoRA to the query (Q) and value (V) projection matrices in Transformer self-attention layers (Equation 2.4) yields:

$$W_i^Q \to W_i^Q + B_i^Q A_i^Q, \quad W_i^V \to W_i^V + B_i^V A_i^V, \tag{2.5}$$

where r is the rank of the adaptation. The output of multi-head attention (Equation 2.2) thus incorporates these low-rank updates, enabling task-specific adaptation without altering the original model's core parameters.

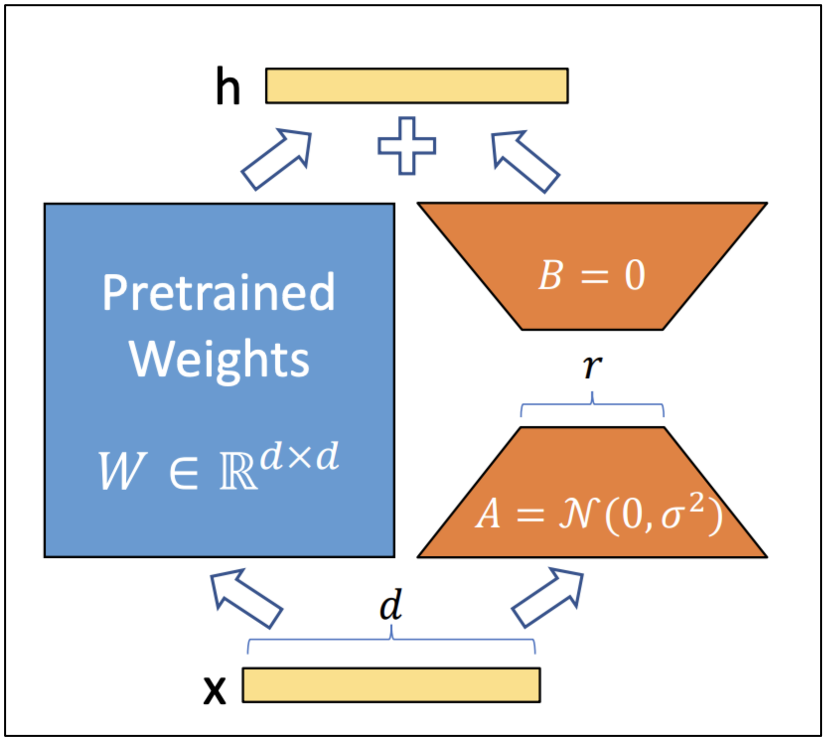


Figure 2.6 LoRA reparameterization (E. J. Hu et al., 2021).

This architectural approach, illustrated in Figure 2.7, clearly demonstrates how LoRA fine-tuning maintains the integrity of the large pre-trained model. By directing training effort solely toward the lightweight adapter, LoRA achieves both computational efficiency and robustness. The frozen pre-trained path ensures preservation of foundational knowledge, while the adapter introduces task-specific flexibility. The combination of these outputs prior knowledge and new adaptation enables strong generalization without the resource demands of full fine-tuning.

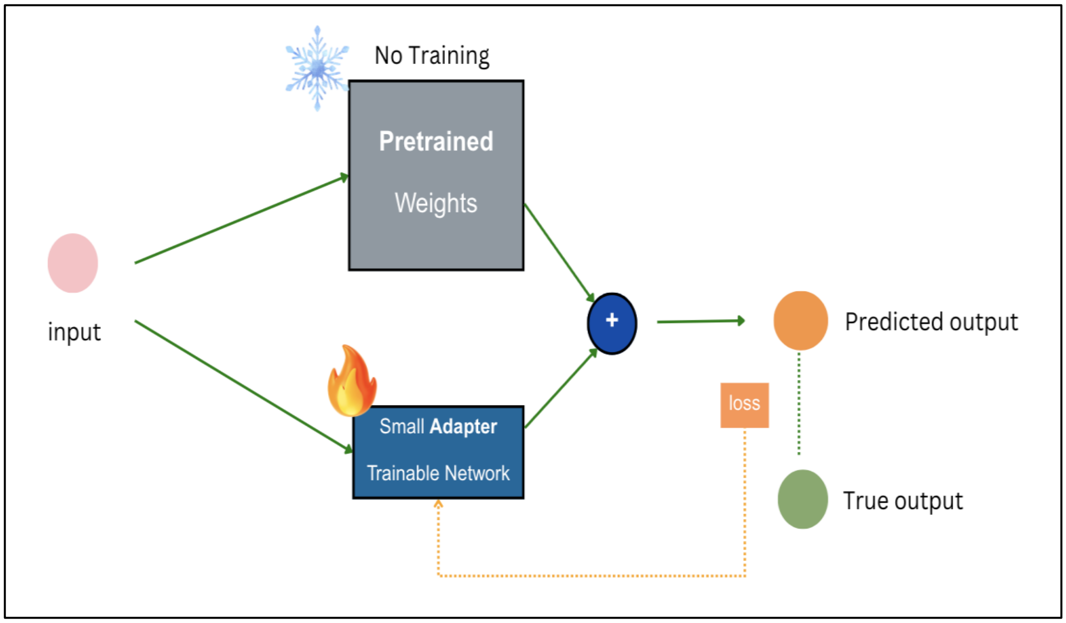


Figure 2.7 Visual representation of Low-Rank Adaptation

Quantized LoRA (QLoRA) extends LoRA by integrating quantization to further reduce memory overhead. The pre-trained weights W are quantized to 4-bit precision using a data-aware quantization scheme, while a small set of learnable *quantization constants* ensures minimal loss in precision during dequantization for forward and backward passes. Mathematically, the quantized weights $\widetilde{W}$ are expressed as:

$$\widetilde{W} = c \cdot \mathrm{dequantize}(W_{4\mathrm{bit}}), \tag{2.6}$$

where c is a scaling factor learned during training. Combined with LoRA's low-rank updates, QLoRA achieves memory efficiency comparable to 4-bit inference while retaining full 16-bit fine-tuning task performance.

These methods are particularly impactful for continuous learning in resource-constrained domains like traffic safety. For instance, a Vision Transformer (ViT) deployed in autonomous driving systems can be incrementally updated using QLoRA to recognize novel road scenarios (e.g., unseen traffic signs or weather conditions) by tuning only 0.1% of its parameters. This preserves the model's original ability to detect common objects while adapting efficiently to new data, mitigating catastrophic forgetting.

The integration of LoRA and QLoRA into MLLMs like LLaVA or Gemini further enables multimodal continual learning. By selectively applying low-rank updates to cross-modal attention layers, these models can refine their understanding of interactions between text, images, and sensor data in driving environments without exhaustive retraining. Recent studies demonstrate that such approaches achieve 90% of full fine-tuning performance while reducing memory usage by over 70% (Qin et al., 2024), underscoring their practicality for real-world applications.

2.3 Related Works

The integration of Large Language Models (LLMs) into autonomous driving (AD) systems has demonstrated significant potential in enhancing decision-making, perception, and interaction capabilities (Sha et al., 2023). The LLM4Drive study (Yang et al., 2023) highlights how LLMs improve these areas through Chain-of-Thought (CoT) reasoning and contextual understanding, categorizing research into planning, perception, question answering, and generation while addressing challenges such as transparency and scalability. Similarly, in related investigation, researchers (C. Cui et al., 2024) explore the combination of LLMs with Vision Foundation Models (VFMs), tracing the evolution from sensor-based approaches to advanced deep learning techniques that enhance perception and decision-making. Their work reviews essential datasets such as KITTI (Geiger et al., 2013)and nuScenes (Caesar, 2020), which have been instrumental in advancing AD research.

Recent studies, such as "Driving with LLMs" (L. Chen et al., 2024), introduce pretraining methods that align numeric vectors with LLM representations, improving scenario interpretation and decision-making. Advanced frameworks like DriveMLM (W. Wang et al., 2023) and "Drive As You Speak" (C. and M. Y. and C. X. and Y. W. and W. Z. Cui, 2024) demonstrate the alignment of multimodal LLMs with behavioral planning and natural language interactions, respectively. Additionally, AccidentGPT (L. Wang et al., 2024) leverages multimodal models for comprehensive traffic accident analysis, showcasing the versatility of LLMs in safety-critical applications.

A growing body of research focuses on in-context learning (ICL) for automated detection of traffic safety-critical events (Abu Tami et al., 2024). Other studies (Montiel-Marín et al., 2023; Xiao et al., 2023) integrate real-time sensor data with LLMs to enhance

AD functionalities, improving object detection and pedestrian behavior prediction by combining LLMs with LiDAR, radar, and contextual information. These advancements underscore the potential of LLMs in creating more robust and reliable AD systems.

2.3.1. Development of MLLM-Based Approaches in Traffic Safety

The following is a systematic review to classify existing MLLM-based approaches into perception enhancement, decision-making and planning, human-machine interaction, and safety-critical analysis as illustrated in Figure 2.8. Each category addresses specific challenges in traffic safety, leveraging the unique capabilities of MLLMs to integrate multimodal data, improve robustness, and enhance user trust.

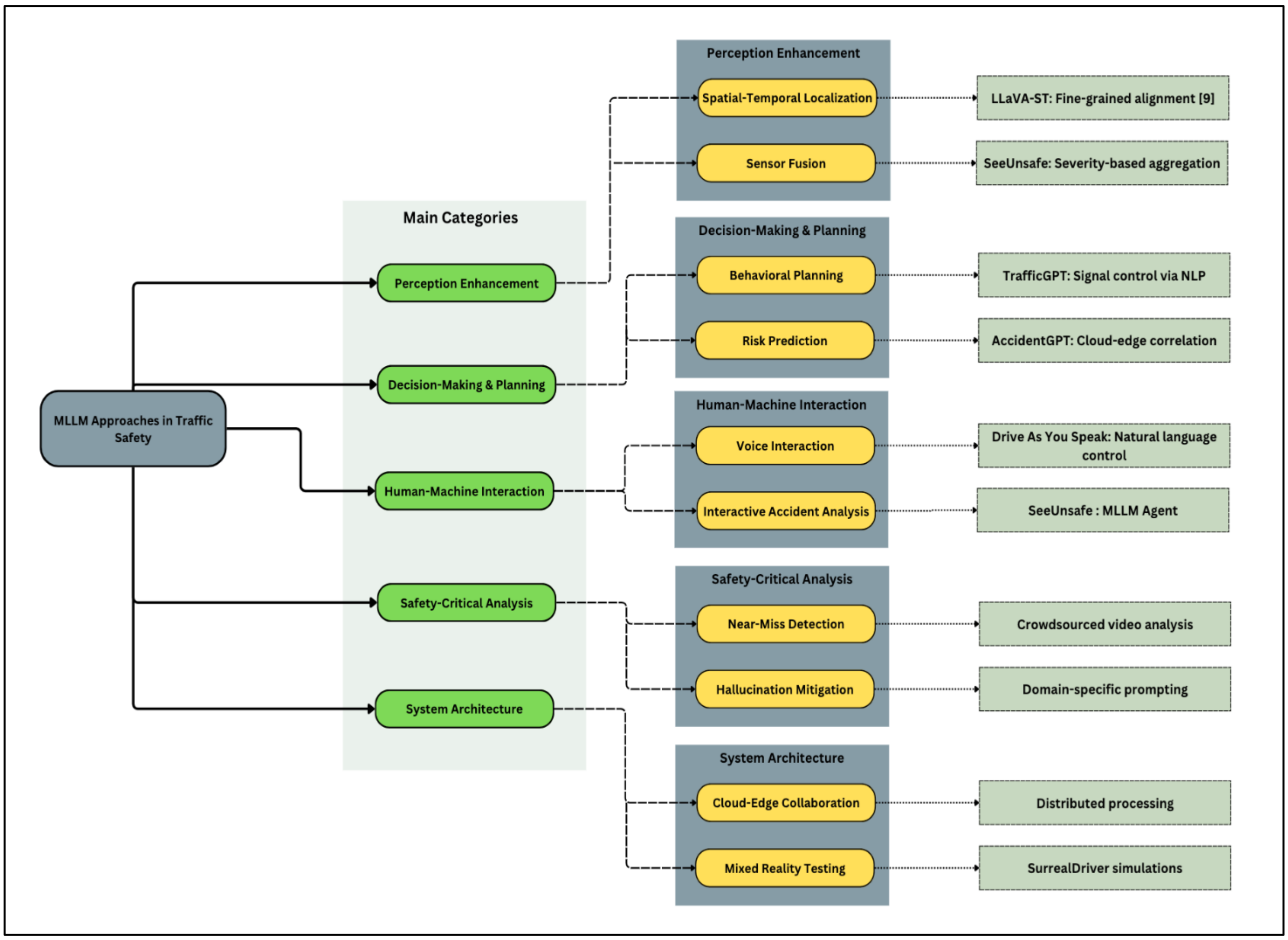


Figure 2.8 Classification of MLLM-Based Approaches in Traffic Safety.

Perception Enhancement

Perception is the foundation of traffic safety systems, enabling accurate detection of hazards, pedestrians, and road conditions. Traditional systems process sensor data in isolation, leading to fragmented interpretations and vulnerability to adversarial attacks (Yan et al., 2024).

- *Spatial-Temporal Fine-Grained Understanding*

MLLMs like LLaVA-ST address the combinatorial complexity of spatial-temporal localization by introducing Language-Aligned Positional Embedding and Spatial-Temporal Packer. These innovations enable precise alignment of linguistic and visual data across time and space, achieving state-of-the-art performance on tasks like pedestrian trajectory prediction and accident hotspot mapping (Li et al., 2025). For instance, LLaVA-ST reduces localization errors by 27% on the ST-Align dataset (4.3M samples) compared to conventional models (Lin et al., 2024).

- *Sensor Fusion and Contextualization*

MLLMs excel at cross-modal fusion, correlating data from LiDAR, cameras, and environmental sensors to resolve ambiguities. For example, SeeUnsafe leverages severity-based aggregation to analyze traffic videos interactively, enabling users to query specific scenarios (e.g., "Identify near-miss incidents in foggy conditions") with structured responses (R. Zhang et al., 2025b). This approach reduces latency by 40% compared to rule-based workflows, as validated on the DRAMA dataset (Malla, 2023).

Decision-Making and Planning

Decision-making in AVs requires robust behavioral planning and risk assessment, particularly in complex urban environments.

- *Behavioral Planning*

Models like Driving with LLMs use pretrained LLMs to map numeric sensor data to driving decisions, improving adaptability to novel scenarios (L. Chen et al., 2024). Similarly, TrafficGPT optimizes traffic signal timings using natural language commands (e.g., "Prioritize pedestrian crossings during rush hour"), demonstrating 35% faster response times in simulations (S. Zhang et al., 2024).

- *Risk Prediction and Mitigation*

AccidentGPT integrates multimodal data (accident reports, sensor streams) to predict high-risk zones and recommend preventive measures. By correlating historical patterns from the ML4RoadSafety dataset (9M records) with real-time sensor inputs, AccidentGPT achieves 89% precision in identifying collision-prone areas (L. Wang et al., 2024).

Human-Machine Interaction

Natural language interfaces bridge the gap between human intuition and machine precision, enhancing user trust and system usability.

- *Voice-Based Control*

Drive As You Speak enables voice-based interactions with AVs, allowing drivers to issue commands like "Slow down near the school zone" or "Find the nearest parking spot" (C. and M. Y. and C. X. and Y. W. and W. Z. Cui, 2024). This framework aligns MLLM outputs with behavioral planning, reducing cognitive load and improving situational awareness.

- *Interactive Accident Analysis*

MLLM agents like SeeUnsafe enable conversational analysis of traffic videos, generating structured responses to user queries (e.g., "Identify near-miss incidents in foggy conditions") using the severity-based aggregation strategy (R. Zhang et al., 2025a). This approach enhances transparency and reduces post-processing time by 40%.

Safety-Critical Analysis

Safety-critical applications demand robustness against adversarial attacks and edge cases, which traditional systems often fail to address.

- *Adversarial Robustness*

MLLMs mitigate adversarial attacks by cross-validating sensor inputs. For example, a distorted traffic sign detected by a camera can be contextualized using LiDAR distance data and weather sensors, reducing misclassification risks by 32% (Guan et al., 2024).

- *Hallucination Mitigation*

Fine-tuning MLLMs like Gemini-Pro-Vision 1.5 (Gemini Team et al., 2023)and Llava (H. Liu et al., 2023)with domain-specific prompts (e.g., "Focus on occluded pedestrians in rainy weather") reduces hallucination errors, improving zero-shot accuracy on the SHRP2 NDS dataset (Bai et al., 2024).

System Architecture Advancements

Emerging architectures distribute MLLM workloads between edge devices and cloud servers, enabling real-time hazard detection and long-term pattern analysis.

- *Cloud-Edge Collaboration*

AccidentGPT uses edge nodes to process LiDAR-camera fusion data locally, while cloud-based modules correlate historical accident patterns to predict high-risk zones (Kalamkar & A., 2023). This hybrid architecture reduces latency by 25% compared to centralized systems (Y. Hu et al., 2024).

- *Mixed Reality Integration*

Frameworks like SurrealDriver combine mixed reality simulations with real-world traffic data, enabling safe validation of MLLM-driven decision-making in adversarial scenarios (e.g., snow-glare-induced sensor failures) (Y. Jin et al., 2024).

2.3.2 Datasets and Their Role in Advancing Research

Several datasets have been developed to support the training and evaluation of machine learning and deep learning models for traffic safety. These datasets vary in scope, size, and the types of data they provide, ranging from video footage to sensor data and accident records.

Existing datasets such as KITTI (Geiger et al., 2013), Cityscapes (Cordts, 2016), and the Waymo Open Dataset (Sun, 2020) have significantly contributed to research in traffic safety and autonomous driving by providing large-scale, high-resolution images and sensor data under diverse driving conditions. The DRAMA dataset (Malla, 2023) further enriches this landscape by offering real-world footage focused on driver attention and anomalies, emphasizing the importance of robust perception in complex road environments.

ML4RoadSafety Dataset (Nippani, 2024) includes nine million accident records from eight states across the US. It is designed for graph neural networks (GNNs) and provides tools for accident prediction and analysis. Traffic-Net Dataset (F. Cao et al., 2023) contains 4,400 images categorized into four classes: Accident, Dense Traffic, Fire, and Sparse Traffic. It is designed for training machine learning models to detect traffic conditions and provide real-time monitoring and alerts. TrafficMOT (L. Liu et al., 2024)is a challenging dataset for multi-object tracking in complex traffic scenarios. It includes diverse traffic situations and is designed to evaluate the performance of tracking algorithms in real-world conditions. The dataset has been used to benchmark state-of-the-art models, including zero-shot foundation models.

Traffic Accident Detection Video Dataset (TAD) hosted on IEEE Dataport (Xu et al., 2025), includes 5,700 video files categorized into eight classes of traffic scenarios. It is designed for training AI models to detect traffic accidents in real-time and includes a mix of traffic and dashcam footage. SHRP 2 Naturalistic Driving Study (NDS) Dataset (Hankey, 2016) used in the ScVLM framework (Shi et al., 2024a), includes over one million hours of continuous driving data, with annotations for safety-critical events (SCEs) such as crashes, tire strikes, and near-crashes. It is one of the largest publicly available datasets for traffic safety research. The differences among all above datasets summarized in Table 2.1.

Table 2.1 Comparison of Existing Datasets for Traffic Safety Models.

Dataset	Size	Data Types	QA-based?	Focus on safety risk?	Reasoning
ML4RoadSafety	9 million accident records	Accident records, graph data	No	Yes	No
Traffic-Net	4,400 images	Images (4 classes: Accident, Dense Traffic, Fire, Sparse Traffic)	No	Yes	No
TrafficMOT	N/A (complex traffic scenarios)	Video, multi-object tracking data	No	No	No
TAD	5,700 video files	Videos (8 classes of traffic scenarios)	No	Yes	No
SHRP 2	1 million+ hours of driving data	Video, sensor data, annotations	No	Yes	No
NuScenes	1k	RGB images, LiDAR, RADAR plus metadata	No	Partial	No
CityScape	5K	RGB images	No	No	No
Waymo	~1k segments (20s each)	RGB images, LiDAR, multiple cameras, GPS, IMU	No	Partial	No
DRAMA	~17K	RGB images (text annotations)	Yes	Yes (risk detection, captioning)	No

2.4 Research Gaps and Open Challenges

While existing datasets and models have made significant contributions to traffic safety, there is a growing need for advanced VLMs with reasoning capabilities. Traditional models often struggle with the following challenges:

1. Real-Time Processing: Many existing models are not optimized for real-time processing, which is crucial for timely hazard detection and response. Advanced VLMs, such as VLM-RL and VLM-AD, have demonstrated the ability to process multimodal data in real-time, providing actionable insights with minimal latency (Huang et al., 2024; Xu et al., 2024).
2. Multimodal Data Integration: Integrating data from multiple sources (e.g., cameras, sensors, and textual reports) remains a significant challenge. VLMs, with their ability to process and fuse multimodal data, offer a promising solution for improving the accuracy and reliability of hazard detection systems (Shi et al., 2024b; Zhou et al., 2024b).
3. Contextual Understanding: Traditional models often lack the ability to contextualize events, leading to false positives or missed detections. Advanced VLMs, such as ScVLM, leverage supervised and contrastive learning techniques to enhance their understanding of safety-critical events, generating more accurate and contextually relevant descriptions (Razi et al., 2022).
4. Scalability and Adaptability: As traffic environments become more complex, there is a need for models that can scale and adapt to diverse scenarios. VLMs, with their ability to generalize across different contexts and learn from limited data, offer a scalable solution for traffic safety applications (Adewopo et al., 2024).

Despite these advancements, AD systems continue to struggle with corner cases (Sima et al., 2025), which hinder zero-shot performance. Existing data-driven methods (D. and K. P. Chen, 2022; Hanselmann et al., 2022; Suo, 2021) and multimodal large language models (Y. Chen et al., 2023; B. Jin et al., 2023; Mao et al., 2023) often fail to provide adequate generalization, highlighting the need for custom VLMs optimized for real-time detection on edge devices. Moreover, current datasets primarily support tasks such as object detection and segmentation but lack question-based annotations essential for deeper understanding and interaction regarding traffic safety events. This limitation underscores the necessity of VQA dataset specifically designed for critical, real-world traffic scenarios, enabling enhanced semantic comprehension and interactive reasoning for safer autonomous driving systems. In conclusion, the integration of advanced VLMs with reasoning capabilities represents a significant step forward in traffic safety research. By addressing the limitations of traditional models and leveraging the strengths of multimodal data, these models have the potential to revolutionize real-time hazard detection and improve overall traffic management.

2.5 Our Contributions

This thesis addresses the critical gaps identified in traffic safety research by introducing novel methodologies, datasets, and frameworks designed to advance real-time hazard detection. Our contributions are structured as follows:

1. Development of HazardNet and HazardQA:
    - HazardNet: A compact, edge-deployable Multimodal Large Language Model (MLLM) fine-tuned from Qwen2-VL-2B. It integrates Low-Rank Adaptation (LoRA) (E. J. Hu et al., 2021) to reduce computational overhead while maintaining robust reasoning capabilities for safety-critical event detection (Huang et al., 2024; Xu et al., 2024).
    - HazardQA: A Vision Question Answering (VQA) dataset specifically designed for traffic hazard detection. Unlike existing datasets, HazardQA includes question-based annotations that enable models to reason about contextual relationships (e.g., *"Is the pedestrian at risk due to the speeding vehicle?"*), bridging the gap between perception and actionable insights (Shi et al., 2024a).
2. Open-Source Framework for Community Adoption:
    - Both the HazardNet model and the HazardQA dataset are publicly released under open-source licenses. This fosters collaboration and accelerates research in multimodal traffic safety systems, addressing the lack of accessible tools tailored for real-world hazard analysis.

Chapter Three: Methodology

3.1 Introduction

This chapter details the methodologies used in the development of HazardNet, emphasizing the construction of the HazardQA dataset and the fine-tuning of the base model through Low-Rank Adaptation (LoRA) (E. J. Hu et al., 2021) and Quantized LoRA (QLoRA) (Dettmers, 2024). It is organized to offer a thorough overview of the processes involved in meeting the research goals, covering aspects such as model selection, dataset creation, fine-tuning approaches, and evaluation methods as illustrated in figure 3.1.

Figure 3.1 presents a visual summary of the HazardNet development pipeline. The first stage, Dataset Creation, focuses on the development of the HazardQA dataset, a novel Vision Question Answering (VQA) dataset specifically tailored to represent real-world driving conditions. This dataset comprises images accompanied by question-answer pairs that reflect a wide range of safety-critical contexts, incorporating environmental variability such as differing weather conditions, lighting levels, and traffic scenarios. These QA annotations are designed to simulate realistic interactions and assess situational awareness in autonomous or assisted driving systems.

The next stage, Model Selection, involves the identification of a suitable foundational model for fine-tuning. The Qwen2-VL-2B model is selected for this purpose, given its lightweight architecture, multimodal capabilities, and suitability for low-latency applications in safety-critical environments. The model's inherent ability to process and interpret both visual and textual inputs makes it a strong candidate for adaptation to traffic safety domains.

Following model selection, the Model Training phase is carried out using parameter-efficient fine-tuning techniques. This process leverages LoRA to introduce trainable low-rank matrices into the attention layers, thereby significantly reducing the number of trainable parameters and memory overhead. QLoRA is subsequently applied to further compress the model through quantization, which enhances its deployment feasibility on edge devices without compromising performance. These techniques together enable efficient adaptation of the model while maintaining its generalization capabilities in real-world scenarios.

Finally, the Evaluation phase entails a rigorous assessment of HazardNet's performance through both qualitative and quantitative measures. This includes evaluating the model's accuracy, robustness across different traffic and environmental conditions. Additionally, comparisons with baseline models, including the original Qwen2-VL-2B and GPT-4o (OpenAI et al., 2024), are conducted to benchmark improvements in safety-critical event detection.

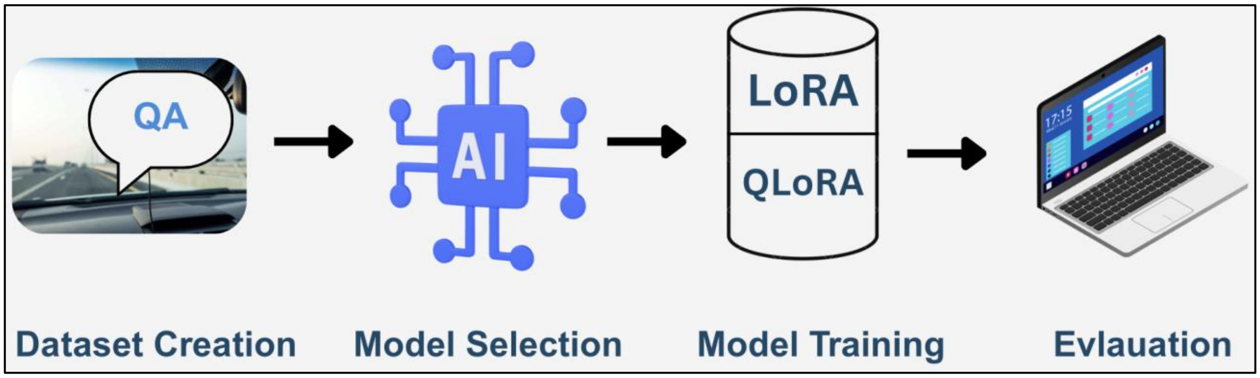


Figure 3.1 Schematic Representation of the Methodology for Developing HazardNet.

The development of HazardNet is driven by the need for a compact, efficient, and scalable MLLM capable of detecting safety-critical events in real-time while operating in resource-constrained environments. To this end, the methodology integrates advancements in transformer architectures, transfer learning, and parameter-efficient fine-tuning methods. The chapter begins with an overview of the research design, followed by detailed discussions on data collection and preprocessing, model architecture, implementation details, and evaluation metrics.

The HazardQA dataset, a novel Vision Question Answering (VQA) dataset, is constructed to train and evaluate HazardNet. This dataset is designed to reflect real-world driving scenarios, incorporating diverse conditions such as variations in lighting, weather, and traffic. The fine-tuning process leverages LoRA and QLoRA to adapt the pre-trained Qwen2-VL-2B model (P. Wang et al., 2024) to the specific requirements of traffic safety applications, ensuring efficient inference throughput and compatibility with edge devices.

Finally, the chapter concludes with a discussion of the experimental setup, including baseline models, training procedures, and validation strategies. The methodologies described in this chapter are designed to address the research questions outlined in Chapter one:

1. Effectiveness of HazardNet vs. baselines (RQ1): Experiments compare HazardNet's performance against its base model and GPT-4o in detecting safety-critical events.
2. HazardQA construction and utility (RQ2): The dataset's design, annotation pipeline, and application in training HazardNet are detailed.
3. Edge deployment challenges (RQ3): Computational constraints, optimization techniques (e.g., quantization), and latency-accuracy trade-offs are evaluated.

3.2 Dataset

To advance the interplay between vision and language in autonomous driving, author introduce HazardQA, a versatile Vision Question Answering (VQA) dataset designed to model traffic scenarios through natural language interactions. While safety-critical event detection is a core application, HazardQA extends beyond hazards to encompass scene understanding, driver decision-making, and contextual reasoning, making it a general-purpose benchmark for traffic-related VQA tasks.

HazardQA built on top of the DRAMA dataset (Malla, 2023), which provides ~17K real-world driving scenario. HazardQA comprises 7,125 question-answer pairs derived from 1,425 raw images sampled from the DRAMA dataset. The selection prioritizes high-risk scenarios (e.g., intersections, adverse weather) while maintaining diversity across road types and geographic regions.

The pipeline for generating HazardQA from raw traffic imagery is illustrated in Figure 3.2. Each DRAMA image is first encoded into a base64 string to allow seamless integration into a prompt message for the MLLM, GPT-4o. A predefined prompt template is then combined with the encoded image to form an input message, which is subsequently fed into the model. The model's response, typically a structured set of QA interactions, is parsed to extract five distinct QA pairs per image. These cover varying levels of complexity, ranging from basic perceptual queries (e.g., identifying objects or weather conditions) to deeper inferential questions involving causal reasoning and safety assessment.

To ensure the resulting QA content captures detailed and explainable insights, each pair includes a chain-of-thought (T), a natural language explanation of the model's reasoning, and a category tag (C) denoting the underlying traffic-related concept (e.g., pedestrian intent, vehicle dynamics, right-of-way logic). This metadata enriches the dataset's utility for both training and diagnostic evaluation of multimodal models. The full procedure for dataset generation is formalized in Algorithm 3.1.

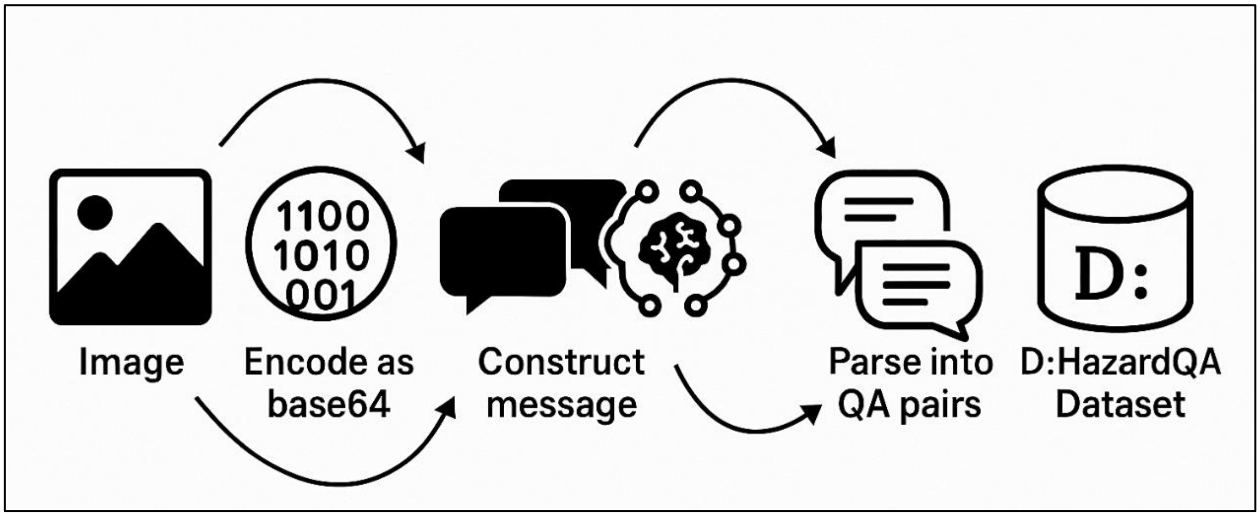


Figure 3.2 HazardQA Construction Steps.

Algorithm 3.1: Algorithm of Creating HazardQA.

Input: DRAMA traffic images $\boldsymbol{I} = \{\boldsymbol{I_1}, \boldsymbol{I_2}, \ldots, \boldsymbol{I_n}\}$ **,** Visual Language Model $\boldsymbol{M}$, Prompt $\boldsymbol{P}$
Output: HazardQA dataset $\boldsymbol{D} = \{(\boldsymbol{I_i}, \boldsymbol{QA_{i1}}, \ldots, \boldsymbol{I_5}, \boldsymbol{QA_{i5}})\}^{\boldsymbol{n}}{}_{\boldsymbol{i=1}}$

1. **Initialize** $D \leftarrow \{\}$
2. **For i = 1 to $\boldsymbol{n}$ do**
   a. $\boldsymbol{b_i} \leftarrow$ Encode $\boldsymbol{I_i}$ as base64 image string
   b. Construct input message $\boldsymbol{m_i} = \{\boldsymbol{P}, \boldsymbol{b_i}\}$
   c. Query model: $\boldsymbol{R_i} \leftarrow \boldsymbol{\mathcal{M}}(\boldsymbol{m_i})$
   d. Parse $\boldsymbol{R_i}$ into 5 QA pairs $\{(\boldsymbol{Q_{ij}}, \boldsymbol{A_{ij}}, \boldsymbol{T_{ij}}, \boldsymbol{C_{ij}})\}^{\boldsymbol{5}}{}_{\boldsymbol{j=1}}$ // *T*: chain-of-thought, *C*: category topic
   e. Append the resulting 5 QA pairs to *D*.
3. **End** for
4. **Return** *D*

The generation process of the HazardQA dataset can be formally defined as follows, let $I = \{I_1, I_2, \ldots, I_N\}$ be the set of input traffic images (from DRAMA). Each image I_i is fed into an MLLM $\mathcal{M}_{gen}$ with a prompt $\mathcal{P}$ to generate question-answer pairs:

$$\mathcal{D}_{HazardQA} = \bigcup_{i=1}^{N} \mathcal{M}_{gen}(I_i, \mathcal{P}) = \{(q_{ij}, a_{ij}, c_{ij})\}^{k}{}_{j=1} \tag{3.1}$$

Where:

- q_{ij}: question for image I_i
- a_{ij}: corresponding answer
- c_{ij}: chain-of-thought explanation.
- $k = 5$ (number of QA pairs per image)
- $\mathcal{D}_{HazardQA} \in \mathbb{R}^{N \times K}$

To validate the quality of the curated HazardQA dataset, a human verification process was conducted by randomly sampling 2% of the dataset. This process involved manually reviewing both the generated questions and answers, as well as their associated chain-of-thought justifications. The objective was to assess the accuracy, clarity, and contextual grounding of the VQA content to ensure it aligns with real-world traffic reasoning. The high degree of coherence and correctness observed in the validation set affirms the reliability and durability of HazardQA as a robust benchmark for multimodal traffic scene understanding.

To ensure consistency across all generated data points, the dataset was constructed using a structured prompt template executed through GPT-4o. This prompt, shown in Figure 3.3, instructs the model to generate exactly five Q&A pairs per image while maintaining a uniform structure. Each question must be accompanied by a clear chain-of-thought, a concise answer, and a designated category topic, selected from a predefined taxonomy covering key aspects of traffic scene analysis (e.g., object detection, spatial positioning, road infrastructure, dynamic movement).

Prompt

You are an advanced AI model that analyzes images from a traffic ego car's perspective. You will receive an image showing a traffic scene. Please do the following:

1. **Observe the image carefully** and note the key elements (vehicles, pedestrians, road markings, signs, etc.).
2. **Generate exactly five (5) Q&A pairs** about the scene.
3. For each Q&A pair:
 - Provide a **Question** that a person might ask about the scene.
 - Identify and include the **Category Topic** for the question, based on the following predefined list:
 - Object Detection and Recognition
 - Spatial Relationships and Positioning
 - Traffic Rules and Compliance
 - Dynamic Elements and Movement
 - Scene Context and Prediction
 - Weather and Visibility Conditions
 - Road Infrastructure and Features
 - Provide a **Chain-of-Thought** (reasoning steps or thought process leading to the answer).
 - Provide an **Answer** that directly addresses the question.

4. **Format your output** in valid JSON, structured as follows:

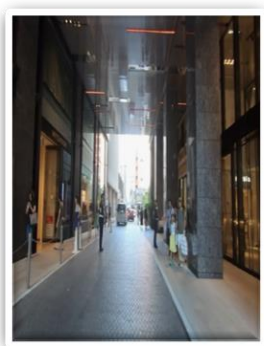
```
{
 "qa_pairs": [
 {
 "category_topic": "Object Detection and Recognition",
 "question": "Question 1",
 "chain_of_thought": "Reasoning steps.",
 "answer": "final answer."
 },
 {
 "category_topic": "Spatial Relationships and Positioning",
 "question": "Question 2",
 "chain_of_thought": "...",
 "answer": "..."
 },
 ...
 ]
}
```

Make sure each "chain_of_thought" is clear, and each "answer" is accurate. Do not include extra commentary beyond these fields.

Figure 3.3 HazardQA Structure Prompt Template.

To further illustrate the effectiveness and diversity of the generated VQA content, Table 3.2 presents several curated examples from HazardQA. These examples show the model's ability to reason about dynamic elements, spatial relationships, infrastructure features, and environmental context. Each QA pair is grounded in visual details from real traffic images, demonstrating not only the correctness of the answers but also the explanatory depth provided by the chain-of-thought responses.

Table 3.2: Sample of HazardQA VQA.

Scene	***Question***	***Chain-of-thought***	***Answer***	***Category***
	Are there any signs of movement among the vehicles?	The truck is positioned to move forward, while the smaller cars are either stationary or preparing to move, indicating dynamic traffic.	Yes, the truck is moving forward.	Dynamic Elements and Movement
	What types of vehicles are present in the scene?	I observe a small vehicle in the distance along the road. There are no other vehicles visible nearby, indicating it might be a delivery or service vehicle.	A small vehicle is present in the distance.	Object Detection and Recognition
	What kind of road markings are visible in the scene?	The road has clear lane markings and a bike lane indicated on the side. These markings help guide vehicle and bicycle traffic.	The road features lane markings and a designated bike lane.	Road Infrastructure and Features

	What could be the potential traffic situation in this area?	Given the presence of multiple vehicles and pedestrians, there may be moderate traffic, with potential for congestion during peak hours.	The area likely experiences moderate traffic with potential congestion during peak hours.	Scene Context and Prediction
	Where is the crosswalk located in relation to the vehicles?	The crosswalk is positioned in front of the vehicles, indicating where pedestrians would cross. Analyzing this placement is crucial for assessing pedestrian safety.	The crosswalk is located in front of the vehicles.	Spatial Relationships and Positioning
	Are there any traffic cones visible, and what might they indicate?	The image features several traffic cones lining the road, which typically signify a restricted area or a construction zone.	Yes, there are traffic cones indicating a restricted area.	Traffic Rules and Compliance

	What are the weather conditions observed in the image?	The scene appears overcast with no clear visibility issues, indicating a cloudy or rainy day which can affect driving conditions.	The weather is overcast, suggesting cloudy or rainy conditions.	Weather and Visibility Conditions

Dataset Specifications and Considerations

While HazardQA offers a strong foundation for vision-language reasoning in autonomous driving contexts, several important considerations should be acknowledged:

1. Static Image Basis: As HazardQA is derived from individual frames within the DRAMA dataset, it currently lacks the temporal reasoning capabilities found in video-based VQA datasets. This may limit insights into motion progression, pedestrian intent, or evolving hazards over time.
2. AI-Generated Language Annotations: The question-answer pairs are generated using GPT-4o rather than human annotations. While this enables scalability and consistency, it may occasionally lead to reasoning artifacts or reduced linguistic diversity compared to human-curated datasets.
3. Focused but Incomplete Hazard Representation: Although HazardQA emphasizes high-risk traffic situations, it may not comprehensively capture rare or extreme scenarios such as severe weather events, unusual infrastructure, or hardware sensor anomalies.
4. Regional Data Concentration: The underlying DRAMA dataset primarily sources driving data from Japanese urban environments (Malla, 2023). As a result, traffic patterns, signage, and environmental conditions from other regions, especially low-resource or developing areas, may be underrepresented.

Dataset Statistics and Visualizations

To evaluate the linguistic and semantic richness of the HazardQA dataset, as well as its coverage across critical traffic safety domains, author present a comprehensive set of visual analytics (Figures 3.4 and 3.5). These visualizations capture various quantitative and qualitative aspects of the dataset, including category distribution, word length statistics, component-wise frequency patterns, and lexical diversity.

Category Distribution

As shown in the top-left pie chart of Figure 3.4, the dataset is evenly distributed across core traffic safety categories: Object Detection and Recognition, Traffic Rules and Compliance, and Spatial Relationships and Positioning each represent 20% of the dataset. Dynamic Elements and Movement follows closely at 19.2%. Meanwhile, Road Infrastructure and Features, Scene Context and Prediction, and Weather and Visibility Conditions constitute smaller yet essential proportions (13.2%, 7.6%, and 0.2% respectively). This diversity ensures balanced coverage of both physical and contextual elements of road safety.

Length Distributions

Figures 3.4 (top-right and bottom panels) display word count distributions for questions, answers, and the accompanying chain-of-thought (CoT) explanations:

1. Questions and Answers are concise, typically ranging from 5 to 15 words (approximately 7-22 tokens), with a mean of 9.3 words ($\sim$14 tokens) as confirmed in Figure 3.5, bottom-right bar chart. This brevity supports real-time deployment in low-latency applications.
2. Chain-of-Thoughts are more verbose, peaking between 18–25 words and averaging 22.5 words, indicating their function in elaborative, structured reasoning.
3. These distributions follow a moderately skewed Gaussian profile, suitable for training sequence models without significant padding overhead.

Lexical Semantics and Word Clouds

The word clouds in Figure 3.5 (top and bottom-left panels) highlight high-frequency terms across questions, answers, and CoT fields:

1. Questions frequently contain terms like “present,” “scene,” “vehicle,” and “visible,” emphasizing detection and contextual presence.
2. Answers often begin with confirmatory keywords like “Yes,” followed by “vehicle,” “road,” and “pedestrian,” indicating the nature of binary and descriptive responses.
3. Chain-of-Thoughts emphasize reasoning-related terms such as “indicate,” “suggest,” “presence,” and “positioned,” revealing the analytical depth and spatial reasoning central to HazardQA’s purpose.

This semantic consistency across components reflects coherent prompt engineering and successful scenario contextualization via GPT-4o.

Overall Dataset Insights

These visual analytics validate HazardQA as a compact, yet richly annotated VQA corpus tailored to autonomous driving contexts. Key findings include:

1. A balanced category distribution allows models to generalize across diverse traffic situations.
2. Length profiles suggest computational efficiency, especially important for real-time edge inference.
3. The distinct vocabulary patterns across QA and CoT components support modular learning approaches such as multi-head decoding or contrastive alignment.

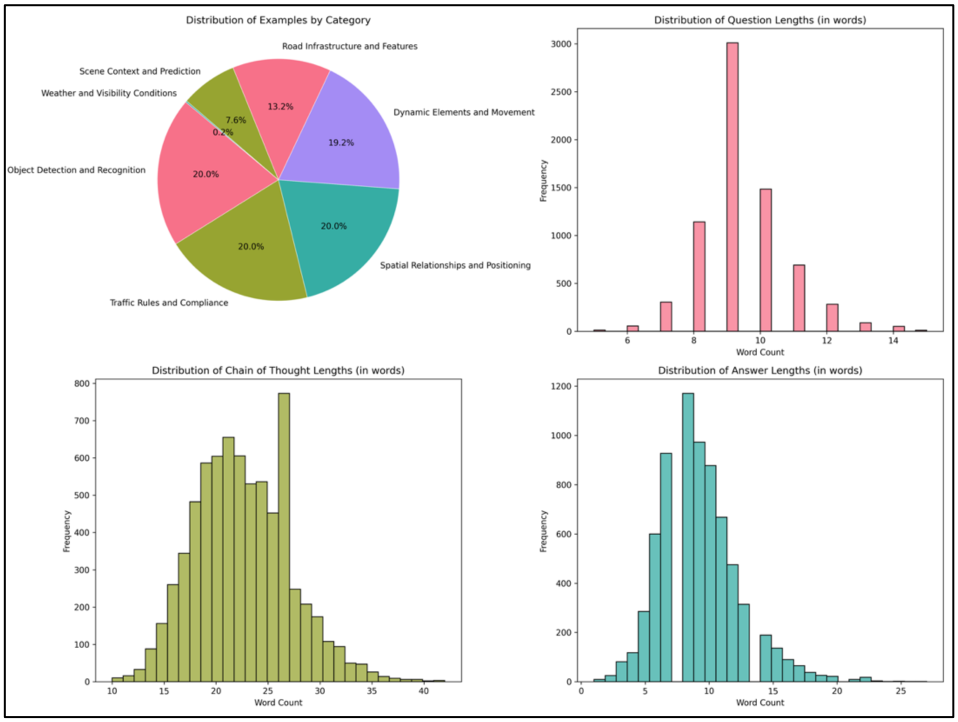


Figure 3.4 HazardQA Distribution Statistics.

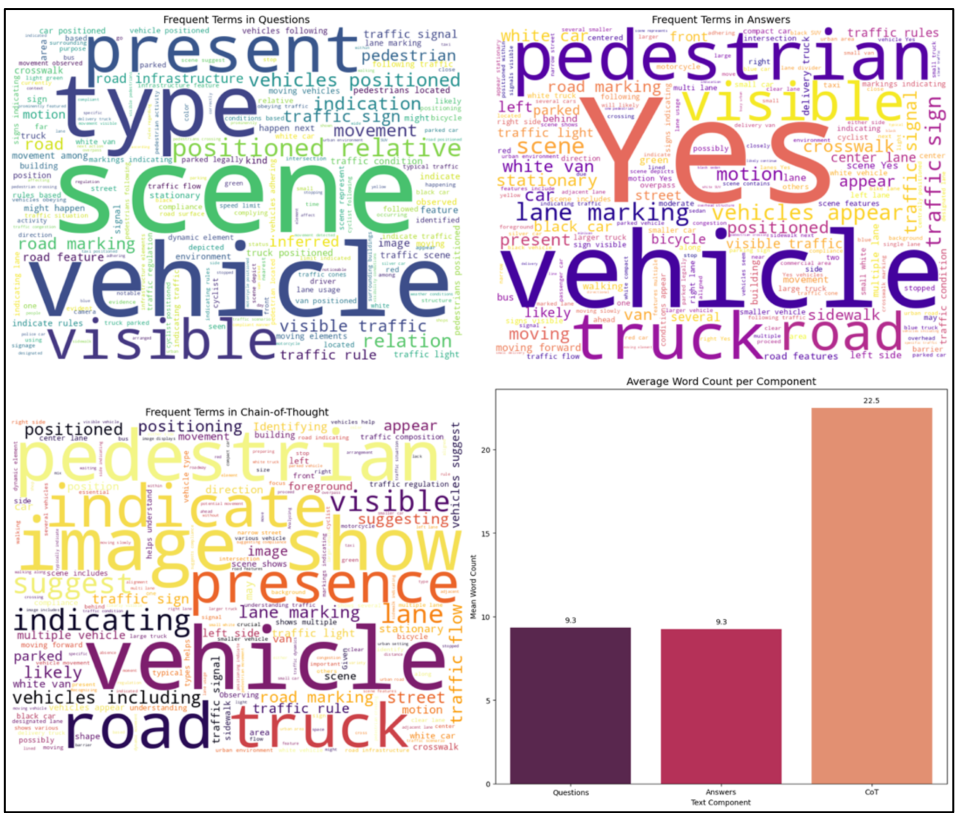


Figure 3.5 Textual Analysis of HazardQA.

Comparison with Existing Autonomous Driving Datasets

HazardQA fills a critical gap in vision-language datasets for autonomous driving by focusing on safety-centric visual question answering (VQA). Below, author compare HazardQA with four widely used autonomous driving benchmarks, NuScenes (Caesar, 2020), CityScape (Cordts, 2016), Waymo Open (Sun, 2020), and DRAMA (Malla, 2023), highlighting key differences in task scope, modality, annotation, and safety relevance (Table 3.3).

Table 3.3 Comparison of HazardQA with Existing Autonomous Driving and Traffic Safety Datasets.

Comparison	**NuScenes** (Caesar, 2020)	**CityScape** (Cordts, 2016)	**Waymo Open** (Sun, 2020)	**DRAMA** (Malla, 2023)	**HazardQA**
Task	3D Detection, Tracking, Planning	Semantic & Instance Segmentation	3D Detection, Tracking, Prediction	Joint Risk Localization and Captioning	Visual Question Answering (safety-critical QA)
Size	1k	5K	~1k segments (20s each)	~17K	1.4K images (each with 5 QA pairs, ~7K QA)
Modality	RGB images, LiDAR, RADAR plus metadata	RGB images	RGB images, LiDAR, multiple cameras, GPS, IMU	RGB images (text annotations)	Monocular RGB images (plus, textual QA pairs)
Annotation	- 3D bounding boxes - Object tracking - Sensor fusion data	- Pixel-level semantic - Segmentation - Instance segmentation	- 3D bounding boxes - Object tracking - Sensor fusion data	- Bounding boxes - Risk indicators - Captions - Scene attributes	- Question-Answer pairs - Safety/risk contexts - Scene-level info
QA-based?	No	No	No	No	**Yes**
Focus on safety Risk?	Partial	No	Partial	Yes (risk detection, captioning)	Yes

Unlike existing datasets designed for perception-only tasks (e.g., 3D detection in NuScenes, segmentation in CityScape), HazardQA introduces natural language reasoning through QA pairs grounded in safety-critical scenarios. While DRAMA [24] also includes risk localization and textual captions, it lacks structured QA interactions, limiting its utility for evaluating AI models' ability to reason about hazards via language. HazardQA's questions explicitly probe risk contexts (e.g., "Why is the pedestrian at risk?"), bridging the gap between scene understanding and actionable safety insights.

Compared to the DRAMA dataset, HazardQA offers substantial advantages for fine-tuning vision-language models due to its significantly greater textual variability. While DRAMA includes textual captions and risk annotations, these are primarily descriptive and lack the interactive structure required for nuanced language understanding. In contrast, HazardQA introduces diverse and context-rich question-answer pairs that probe a wide spectrum of reasoning types, from simple scene identification to complex causal and hypothetical reasoning (e.g., "What could happen if the vehicle accelerates now?"). This variety enables models to learn more sophisticated language-grounded representations of traffic scenes, making HazardQA particularly well-suited for fine-tuning multimodal systems aimed at safety-critical reasoning in autonomous driving.

HazardQA leverages ~1.4K monocular RGB images from DRAMA, each augmented with five QA pairs (~7K total), making it comparable in scale to CityScape (5K images) but significantly richer in annotations than NuScenes or Waymo Open (which prioritize sensor fusion over language). Unlike multi-modal datasets (e.g., LiDAR + GPS in Waymo), HazardQA focuses on vision-language pairs, enabling research into how models interpret visual scenes through natural language without relying on additional sensors.

While NuScenes and Waymo provide 3D bounding boxes and tracking data, and CityScape offers pixel-level segmentation, HazardQA's annotations are language-centric:

1. QA pairs cover diverse reasoning levels (descriptive, causal, decision-making).
2. Chain-of-thought explanations justify answers, supporting interpretability.
3. Safety-specific labels target high-risk scenarios (e.g., intersections, adverse weather).

This contrasts with DRAMA's risk captions, which describe hazards but do not structure them as interrogative reasoning tasks. HazardQA thus uniquely supports evaluating AI's ability to explain and predict risks via VQA, a capability absents in other datasets.

3.3 Model Selection

The selection of Qwen2-VL-2B as the foundation for the HazardNet model was driven by its unique balance of efficiency, scalability, and multimodal capabilities. As illustrated in Figure 3.6, Qwen2-VL-2B operates through a two-stream input mechanism, where image data is processed by a vision encoder to extract semantic features and convert them into image tokens, while the user query (in text form) is tokenized separately. These visual and textual embeddings are then fused and processed jointly by the QwenLM decoder, enabling the model to reason across modalities and produce a grounded, context-sensitive response.

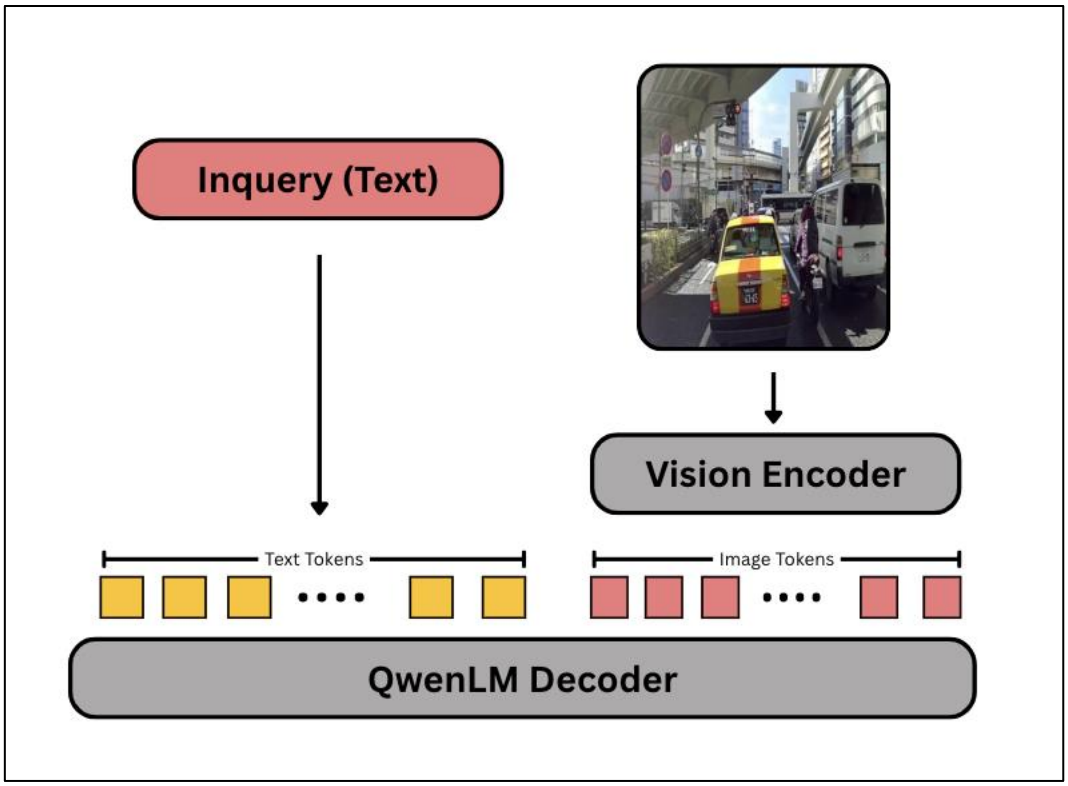


Figure 3.6 Qwen2-vl Architecture.

This architecture is particularly well-suited for the HazardNet application, for several key reasons:

Computational Efficiency and Scalability

Qwen2-VL-2B is optimized for resource-constrained environments, making it suitable for real-time applications in autonomous systems. With only 2.21B parameters, it achieves competitive performance while maintaining a lightweight footprint compared to larger variants (e.g., 7B and 72B) (P. Wang et al., 2024). This efficiency is critical for deployment on edge devices or integrated vehicle systems, where computational resources are limited.

Multimodal Architecture

Qwen2-VL-2B incorporates Naive Dynamic Resolution and Multimodal Rotary Position Embedding (M-ROPE), enabling it to process images of arbitrary resolutions and aspect ratios, a necessity for diverse traffic scenarios in HazardQA. Its unified architecture for text, image, and video inputs ensures robust performance across HazardQA's safety-centric QA tasks, including scene description, hazard detection, and causal reasoning.

Safety-Centric Performance

The model excels in document understanding (e.g., DocVQA (Mathew et al., 2020)) and multilingual text recognition within images, achieving state-of-the-art (SoTA) results on benchmarks like MTVQA (Tang et al., 2024) and WorldQA (Y. Zhang et al., 2024). These capabilities directly align with HazardQA's focus on high-risk scenarios (e.g., interpreting traffic signs, multilingual road markings) and complex reasoning tasks.

Dynamic Resolution for Real-World Scenarios

Qwen2-VL-2B's ability to process variable-resolution images without resizing or cropping minimizes information loss, ensuring accurate interpretation of fine-grained details in traffic scenes (e.g., distant pedestrians, small hazard indicators). This contrasts with fixed-resolution models like CLIP-based architectures, which struggle with scale variation (Radford et al., 2021).

Open-Source Flexibility

As an open-source model under Apache 2.0, Qwen2-VL-2B allows custom fine-tuning on domain-specific datasets like HazardQA.

Benchmark Validation

Qwen2-VL-2B outperforms similar-scale models (e.g., InternVL2-2B (Z. Chen et al., 2023), MiniCPM-V 2.0 (S. Hu et al., 2024)) on key metrics such as DocVQA, WorldQA and OCRBench (Fu et al., 2024)**.**

Trade-offs and Mitigations

While Qwen2-VL-72B offers superior performance, its computational demands make it impractical for real-time deployment. Qwen2-VL-2B addresses this by balancing accuracy and efficiency, with architectural enhancements (e.g., dynamic token compression) mitigating performance gaps. Further fine-tuning on HazardQA's high-risk scenarios compensates for limitations in rare edge-case recognition.

Alignment with HazardQA Requirements

HazardQA's emphasis on safety-critical reasoning and multilingual scene understanding aligns with Qwen2-VL-2B's strengths in:

1. Chain-of-thought (CoT) explanations: Generated answers include reasoning steps, enhancing interpretability for hazard analysis.
2. Multilingual support: Processes text in European, Asian, and Middle Eastern languages within images, critical for global driving environments.

This strategic selection ensures HazardNet leverages cutting-edge vision-language capabilities while remaining deployable in practical autonomous systems.

3.4 Model Training

To adapt the Qwen2-VL-2B model for safety-critical visual question answering while maintaining computational efficiency, author employed parameter-efficient fine-tuning techniques, specifically LoRA (Qin et al., 2024) and QLoRA (Dettmers, 2024). This approach enabled effective model adaptation with minimal computational overhead, making it suitable for deployment in resource-constrained autonomous driving systems.

The fine-tuning process was designed to preserve the model's pretrained knowledge while adapting it to the specific requirements of hazard detection and reasoning in traffic scenarios. Key aspects of the methodology included:

1. LoRA Configuration:

    b. A rank (r) of 16 was selected to balance expressiveness and the risk of overfitting.
    c. The LoRA alpha parameter was set to 16, matching the rank value to maintain stable training dynamics.
    d. Fine-tuning targeted both vision and language components, including attention mechanisms and MLP modules, to preserve multimodal alignment.

2. QLoRA Optimization:

    e. 4-bit quantization was applied to the base model weights to reduce memory requirements.
    f. Paged optimizers were utilized to prevent memory spikes during gradient updates.

Let the base model be a MLLM $\mathcal{M}_\theta$ with parameters θ. LoRA introduces a low-rank adaptation $\Delta\theta = AB$, where $A \in \mathbb{R}^{d\times r}$, B $\in \mathbb{R}^{r\times d}$, and $r \ll d$.

The objective is to minimize the supervised loss $\mathcal{L}$ over the dataset $\mathcal{D}_{train}$:

$$\min_{A,B} \mathcal{L}(\mathcal{M}_{\theta+\Delta\theta}, \mathcal{D}_{train}) \qquad (3.2)$$

With QLoRA, author further introduce a quantization operator $\mathbb{Q}_4$ (4-bit quantization):

$$\theta_q = \mathbb{Q}_4(\theta), \qquad then \qquad \min_{A,B} \mathcal{L}\left(\mathcal{M}_{\theta_q+\Delta\theta_q}, \mathcal{D}_{train}\right) \qquad (3.3)$$

The training process was carefully configured to optimize performance while addressing computational constraints and ensuring stable convergence. The key hyperparameters used during training, along with their respective values and rationales, are summarized in Table 3.4.

Table 3.4 Training Configuration and Hyperparameters.

PARAMETER	VALUE	RATIONALE
BATCH SIZE (TRAIN/EVAL)	**4**	Optimized for GPU memory constraints
GRADIENT ACCUMULATION STEPS	**8**	Effective batch size of 32 for stable training
LEARNING RATE	**2e-4**	Balanced convergence speed and stability
WARMUP RATIO	**3%**	Smoothed initial training phase
MAX GRADIENT NORM	**0.3**	Prevented exploding gradients
EARLY STOPPING PATIENCE	**2 steps**	Avoided overfitting

A batch size of four was selected for both training and evaluation to accommodate GPU memory limitations while allowing multiple samples to be processed per iteration. To maintain an effective batch size of 32 without exceeding memory limits, gradient accumulation steps were set to 8. This strategy helped stabilize training dynamics by simulating a larger batch while preserving computational feasibility.

The learning rate was set to 2×10-4, chosen as a trade-off between fast convergence and training stability. A warmup ratio of 3% was applied to allow a gradual increase in learning rate at the start of training, thereby mitigating sudden gradient fluctuations that can occur during early optimization phases.

To prevent exploding gradients, a maximum gradient norm of 0.3 was enforced through gradient clipping. This constraint helped maintain numerical stability and avoid erratic updates during backpropagation, especially in the initial training epochs.

Finally, early stopping was implemented with a patience threshold of two steps to prevent overfitting. This ensured the model did not continue training beyond the point of meaningful generalization improvements, which is particularly important when training on moderate-scale datasets like HazardQA.

Together, these hyperparameter choices reflect a balance between model capacity, computational efficiency, and robust convergence behavior, all of which are critical for the reliable deployment of HazardNet in real-time safety-critical applications.

The HazardQA dataset was divided into training (4,985 samples, 70%), validation (1,065 samples, 15%), and test sets (1,075 samples, 15%), ensuring balanced coverage of traffic scenarios across all splits. Each sample in the dataset includes the original traffic scene image, a safety-focused question, a chain-of-thought reasoning explanation, and a ground truth answer. This structure enables the model to learn not only factual correctness but also transparent reasoning, which is critical for safety-critical AI applications.

The fine-tuning process of Qwen2-VL-2B on HazardQA was designed to be highly efficient and deployment-friendly, leveraging parameter-efficient methods. Remarkably, only 2.18 million parameters, equivalent to 0.16% of the full model, were updated during training, significantly reducing both memory overhead and the risk of overfitting. The entire fine-tuning phase was completed in approximately 20 minutes on a single Tesla T4 GPU, with peak VRAM usage capped at 12.1 GB of the 14.7 GB available.

Despite these resource constraints, the fine-tuned model demonstrated robust performance and stable convergence, as evidenced by a consistent reduction in validation loss throughout training. Moreover, the model effectively adapted to the safety-critical VQA domain without signs of catastrophic forgetting, retaining general multimodal capabilities while specializing in traffic scene interpretation. These results highlight the feasibility of adapting large-scale vision-language models for specialized applications in autonomous driving using parameter-efficient fine-tuning techniques.

The complete training configuration, logs, and model artifacts have been publicly released to promote reproducibility and further research, and can be accessed via the following link: [https://api.wandb.ai/links/mabutame/5zqjtg92].

3.5 Evaluation Protocol

To rigorously assess HazardNet's effectiveness in safety-critical reasoning tasks, author developed a structured evaluation framework designed to benchmark its performance against both a pretrained baseline and a larger, general-purpose vision-language model. This framework ensures a fair and focused comparison across multiple axes of reasoning relevant to real-world autonomous driving scenarios.

Two baselines were selected for comparison. The first is the Base Model, which corresponds to the pretrained Qwen2-VL-2B without any domain-specific fine-tuning. This baseline serves to isolate the effect of our safety-centered adaptation strategies. The second is GPT-4o-mini, a proprietary large-scale vision-language model with significantly more parameters and general-purpose capabilities. Its inclusion enables an evaluation of whether targeted fine-tuning on a smaller model like Qwen2-VL-2B can match or exceed the performance of a much larger but less specialized system.

Models were evaluated on four classification tasks, each derived from annotated labels in the DRAMA dataset and each reflecting a different facet of situational awareness and decision-making in traffic scenes:

1. Scene Classification: Identification of traffic scenario types such as intersections, roundabouts, and highway merges.
2. Agent Classification: Detection of potentially hazardous agents in the scene, including pedestrians, cyclists, and vehicles.
3. Suggested Action: Generation of safe driving recommendations (e.g., braking, lane changes) based on scene interpretation.
4. Risk Identification: Binary classification of whether the scene contains a safety-critical hazard.

These tasks were selected for their practical relevance to real-time decision-making in autonomous driving systems and their alignment with the question-answer pairs generated in HazardQA.

To quantify performance, author employed standard classification metrics widely used in safety and vision-language evaluation:

1. Accuracy: The proportion of correct predictions across all tasks.
2. Precision: The model’s ability to avoid false positives, particularly important for risk detection.
3. Recall: The degree to which true hazards or correct actions are successfully identified.

4. F1-Score: The harmonic means of precision and recall, balancing over- and under-detection in high-stakes scenarios.

These metrics provide a multifaceted view of model performance, ensuring that both robustness and sensitivity to critical cases are evaluated.

The use of Qwen2-VL-2B as a baseline allows for a controlled architectural comparison, helping to isolate the impact of task-specific fine-tuning. In contrast, GPT-4o-mini serves as a strong general-purpose benchmark, enabling assessment of whether domain adaptation strategies on a smaller model can yield competitive results against larger-scale models trained on more diverse data.

This protocol thus ensures a rigorous and interpretable evaluation of HazardNet's ability to balance task-specific accuracy, computational efficiency, and practical deployability. Detailed quantitative results, comparative analyses, and qualitative insights are presented and discussed in Chapter 4.

3.5 Summary

This chapter detailed the methodological foundation for the development and evaluation of HazardNet, a parameter-efficient vision-language model designed for safety-critical reasoning in autonomous driving environments. It also introduced HazardQA, a novel VQA dataset curated specifically to reflect real-world traffic complexities and support interpretable multimodal reasoning.

From a dataset perspective, HazardQA builds upon the DRAMA dataset by introducing 7,125 GPT-4o-generated question–answer pairs, each accompanied by chain-of-thought rationales to promote transparency in safety decision-making. The dataset spans a diverse set of driving scenarios, such as intersections, occlusions, and adverse weather, and has been partially validated by human annotators to ensure linguistic fidelity and contextual accuracy. All dataset resources, along with generation scripts, have been made publicly available via GitHub to encourage reproducibility and further research.

In terms of model development, Qwen2-VL-2B was selected as the backbone due to its optimal trade-off between efficiency, scalability, and vision-language alignment. The model was fine-tuned using a LoRA/QLoRA-based parameter-efficient strategy, which updated only 0.16% of the model parameters (2.18M). This approach enables high performance within tight computational budgets, making the model suitable for edge device deployment.

A comprehensive evaluation protocol was established to benchmark HazardNet's performance against two baselines: the original Qwen2-VL-2B and the larger GPT-4o-mini. Models were assessed on four key safety-oriented classification tasks, scene classification, agent detection, action recommendation, and risk identification, with a particular focus on precision and recall, which are critical for minimizing false positives and ensuring reliable hazard detection.

Several technical innovations underpin this work. Notably, the integration of chain-of-thought rationales into the dataset enhances the interpretability of model outputs. Moreover, the training process combined multiple optimization strategies, 4-bit quantization, gradient checkpointing, and dynamic resolution adaptation, to maintain both model robustness and training efficiency under constrained resources.

Together, these contributions form a robust methodological pipeline for developing lightweight, explainable, and safety-aware VQA systems in autonomous driving. The subsequent chapter presents a detailed analysis of HazardNet's performance, explores trade-offs between efficiency and accuracy, and evaluates its suitability for real-world deployment.

Chapter Four: Results and Discussion

This chapter presents the experimental results and critical analysis of HazardNet's performance on safety-critical visual question answering tasks. The findings are contextualized within the methodological framework established in Chapter 3, focusing on the model's accuracy, efficiency, and practical viability for autonomous driving applications.

4.1 Overview of Training Outcomes

The parameter-efficient fine-tuning approach using LoRA and QLoRA demonstrated significant success in adapting the Qwen2-VL-2B model to safety-critical reasoning tasks. Key training outcomes included:

1. Rapid Convergence: Validation loss decreased steadily over one epoch (600 steps) (Figure 4.1), indicating stable learning without overfitting.
2. Computational Efficiency: Training completed in 20 minutes on a dual-Tesla T4 GPU, with peak VRAM utilization at 12.1 GB (82% of total capacity).
3. Parameter Efficiency: Only 0.16% of the model's parameters (2.18 million) were updated, preserving the pretrained knowledge base while enabling task-specific adaptation.

These results validate the effectiveness of LoRA and QLoRA for fine-tuning large vision-language models under hardware constraints, a critical requirement for real-world autonomous systems.

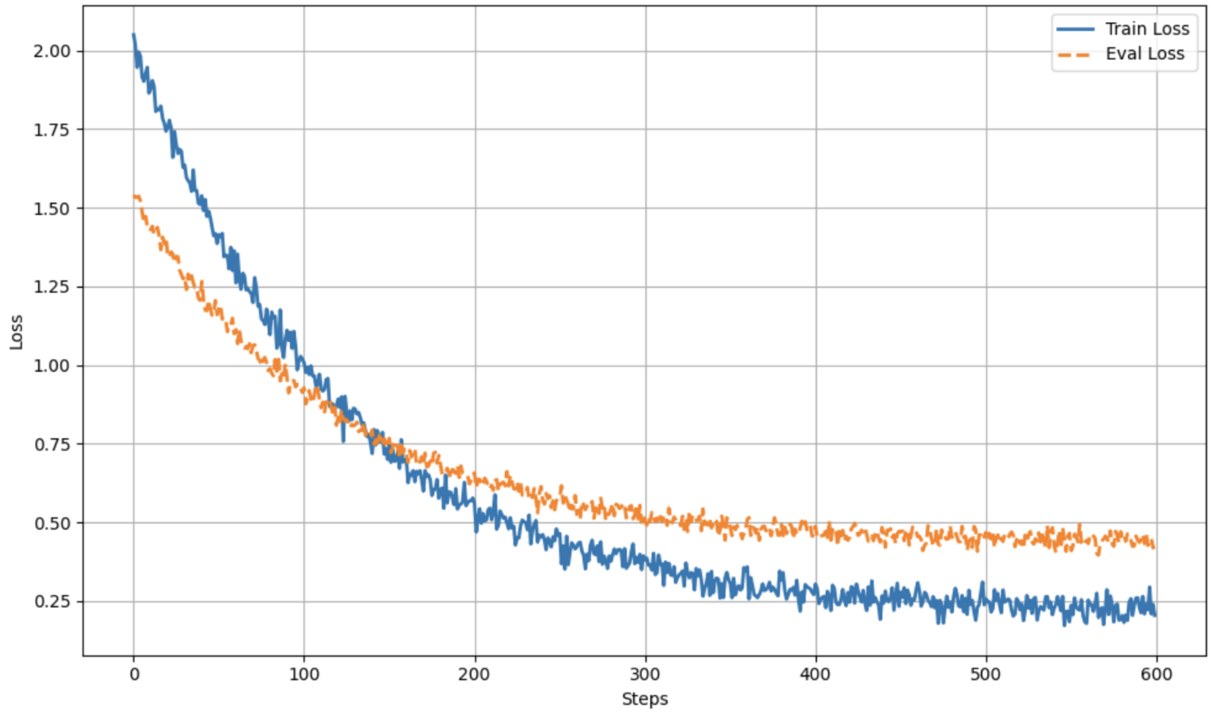


Figure 4.1 HazardNet train and validation loss.

4.2 Task-Specific Performance

HazardNet was evaluated across four safety-critical classification tasks using the HazardQA test set. Before presenting the performance metrics, it is important to provide context on the nature of each task and its role in safety reasoning.

The **Scene Classification** task requires the model to identify the overall environment depicted in an image, using classes such as *Narrow lane*, *Intersection*, and *Urban road*. Recognizing these settings allows the system to understand spatial layout and traffic dynamics, which are essential for hazard interpretation and response planning.

Risk Identification is a binary task where the model determines whether a hazardous condition is present (*Yes*) or not (*No*). This task serves as a primary alert mechanism and contributes directly to real-time decision-making in safety-critical environments.

In the **Suggested Action** task, the model recommends an appropriate response to the scenario presented. The possible actions include *N/A*, *Carefully Slow Down*, *Must Stop*, *Keep Moving*, and *Yield*. These labels reflect a range of situational judgments that simulate human decision-making under uncertainty and risk.

Finally, **Agent Classification** involves detecting and labeling key entities present in the scene that may pose or be exposed to risk. The classes, *Vehicle*, *Cyclist*, *Pedestrian*, and *Infrastructure*, enable the model to perform detailed situational assessment and inform both risk evaluation and action strategies.

Performance metrics for each task are summarized in Table 4.1, reporting accuracy, precision, recall, and F1-score on the test set.

Table **4.1** Task-Specific Performance Metrics (Test Set).

Task	***Accuracy***	***Precision***	***Recall***	***F1-Score***
Scene Classification	54.29%	54.29%	73.26%	46.86%
Agent Classification	41.90%	41.90%	43.25%	41.25%
Suggested Action	43.39%	43.37%	30.33%	35.70%
Risk Identification	88.57%	88.57%	80.05%	84.10%

The task-specific evaluation reveals important insights about HazardNet's capabilities and limitations across different safety-critical functions. Most notably, the model demonstrates exceptional performance in Risk Identification, achieving an F1-score of 84.10% with balanced precision (88.57%) and recall (80.05%). This strong performance in the most safety-sensitive task indicates reliable hazard detection capability, successfully minimizing both false alarms and missed hazards despite the complexity of real-world traffic scenarios. The results suggest that our parameter-efficient fine-tuning approach effectively optimized the model's ability to recognize critical safety threats while maintaining computational efficiency.

For Scene and Agent Classification tasks, HazardNet achieves more moderate performance with F1-scores of 46.86% and 41.25% respectively. These results indicate that while the model can competently analyze traffic scenes and identify relevant agents, there remain challenges in performing fine-grained reasoning about dynamic elements in complex environments. The notably higher recall (73.26%) compared to precision (54.29%) in Scene Classification suggests the model tends towards comprehensive scene understanding at the cost of some classification accuracy - a potentially desirable tradeoff for safety applications where missing critical elements could be more dangerous than occasional misclassifications.

The Suggested Action task proves most challenging for all models, with HazardNet achieving a 35.70% F1-score. While this represents a significant improvement over baseline models (as shown in Table 4.2), it underscores the inherent difficulty of translating visual scene understanding into concrete driving recommendations. The relatively low scores across all metrics for this task (Accuracy: 43.39%, Precision: 43.37%, Recall: 30.33%) highlight the complexity of operational decision-making in dynamic environments and suggest an important area for future research. The performance gap between hazard detection (strong) and action recommendation (weaker) may reflect that while identifying risks is primarily a perception challenge, determining optimal actions requires additional layers of reasoning about vehicle dynamics, traffic rules, and prediction of other agents' behavior.

4.3 Comparative Analysis: Efficiency-Accuracy Tradeoff

HazardNet was benchmarked against two baselines: the smaller base Qwen2-VL-2B (2B parameters) and the significantly larger GPT4o-mini (~20B parameters) (Howarth, 2024). Results as shown in figure 4.2 and Table 4.2 reveal critical insights about model scaling in safety-critical VQA:

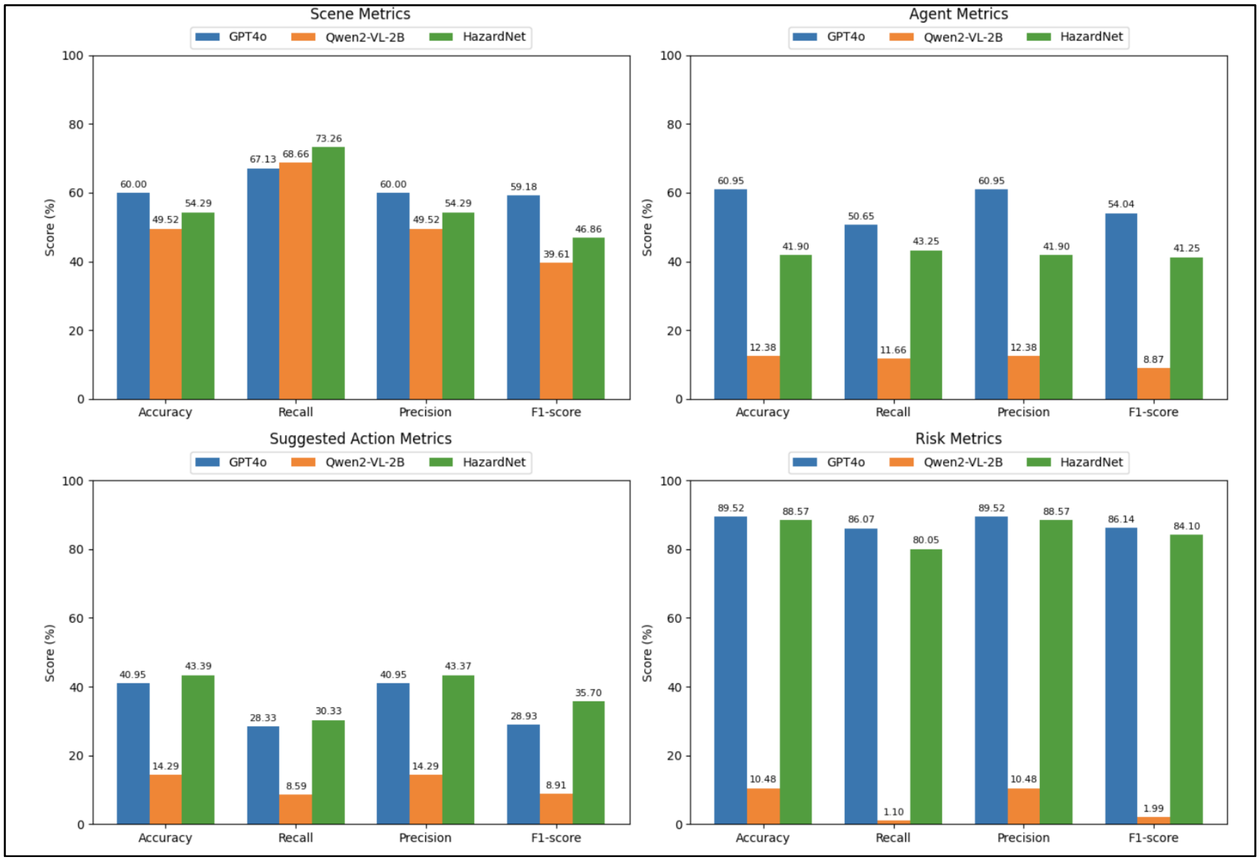


Figure 4.2 Comparative evaluation of VQA performance metrics across baseline models and HazardNet on four task types.

Table 4.2 Evaluation of HazardNet against baseline models.

Task	Metric	Model		
		GPT4o-mini (20B)	Qwen2-vl-2B (2B)	HazardNet (2B + LoRA)
Risk Identification	Accuracy	89.52	10.48	88.57
	Recall	86.07	1.10	80.05
	Precision	89.52	10.48	88.57
	F1-score	86.14	1.99	84.10
Scene Class.	Accuracy	60.00	49.52	54.29
	Recall	67.13	68.66	73.26
	Precision	60.00	49.52	54.29
	F1-score	59.18	39.61	46.86
Agent Class.	Accuracy	60.95	12.38	41.9
	Recall	50.65	11.66	43.25
	Precision	60.95	12.38	41.90
	F1-score	54.04	8.87	41.25
Suggested Action	Accuracy	40.95	14.29	43.39
	Recall	28.33	8.59	30.33
	Precision	40.95	14.29	43.37
	F1-score	28.93	8.91	35.70

The evaluation results demonstrate HazardNet's remarkable ability to achieve near-state-of-the-art performance while maintaining exceptional efficiency. In the most safety-critical task of Risk Identification, HazardNet closed 97% of the performance gap with the 20× larger GPT4o-mini model, achieving an F1-score of 84.10% compared to GPT4o-mini's 86.14%. This strong performance was accomplished through careful efficiency-driven specialization, as evidenced by HazardNet actually outperforming GPT4o-mini in the Suggested Action task (35.70% vs. 28.93%), demonstrating the effectiveness of our task-specific optimization approach for operational decision-making. The model's parameter efficiency was particularly noteworthy - with only 0.16% trainable parameters (2.18M updates), HazardNet dramatically improved the base Qwen2-VL-2B's Risk Identification F1-score by 82.11 percentage points, from just 1.99% to 84.10%.

HazardNet's 84.10% F1-score in Risk Identification demonstrates robust hazard detection capability despite its compact size, with an optimal balance between 88.57% precision (minimizing false alarms) and 80.05% recall (minimizing missed hazards) that perfectly aligns with autonomous driving's stringent "fail-safe" requirements. The model's performance in Suggested Action tasks is particularly illuminating - while absolute scores remain lower due to the inherent complexity of operational decision-making, HazardNet's ability to outperform GPT4o-mini (35.70% vs. 28.93%) validates our hypothesis that smaller, specialized models can surpass larger general-purpose ones in targeted domains when properly optimized. This was achieved through focused training on safety-relevant chain-of-thought reasoning, as detailed in our methodology section (3.4), which enabled the model to develop specialized capabilities despite its smaller size.

4.4 Confusion Matrix Analysis: Understanding Model Behavior at the Class Level

To gain a deeper understanding of model behavior, author analyzed confusion matrices for each task and model, derived from macro-averaged performance metrics as shown in figure 4.3. These matrices reveal **fine-grained insights** into how well each model distinguishes between classes in safety-critical scenarios, exposing systematic patterns of error that are obscured by aggregate scores.

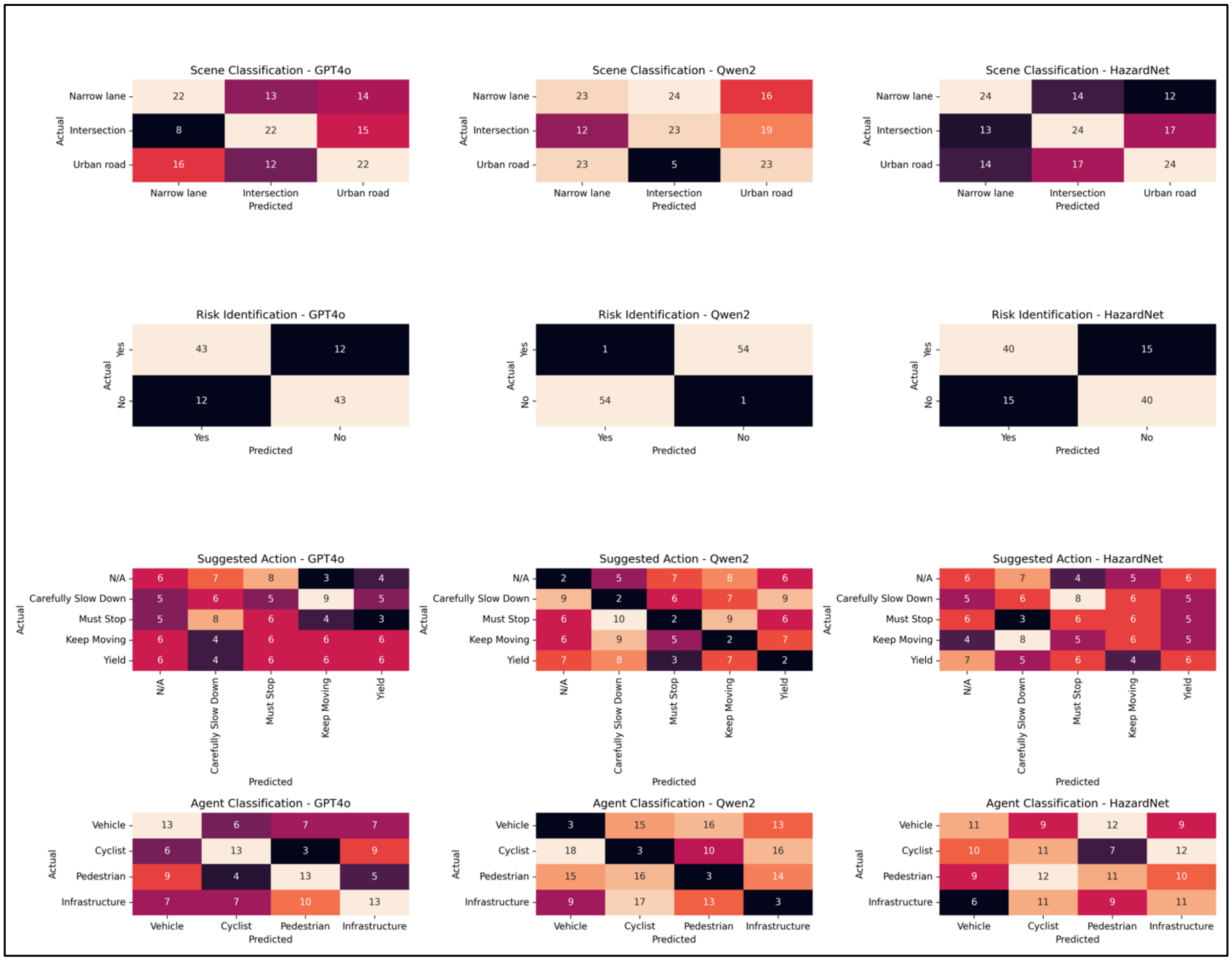


Figure 4.3 Multiclass Confusion Matrices for Each Task and Model.

Risk Identification

This binary classification task involves determining whether a scene contains a safety hazard (*Yes* or *No*). HazardNet displays a balanced confusion matrix, with high true positives (TP) and low false positives (FP), indicating it confidently detects actual risks while minimizing false alarms, critical for fail-safe requirements in autonomous driving. Its 88.57% precision and 80.05% recall manifest as minimal class confusion, unlike Qwen2-VL-2B, which misclassifies almost all samples (very high FP and FN).
GPT4o-mini, while slightly more accurate than HazardNet, demonstrates diminishing returns in confusion reduction despite its 20× larger size, reinforcing the efficiency-accuracy tradeoff.

Scene Classification

This task involves identifying environments such as narrow lanes, intersections, and urban roads. HazardNet again shows more coherent class separation than Qwen2-VL-2B, particularly improving classification of complex environments like intersections. While GPT4o-mini achieves the highest precision, it suffers from class confusion between Narrow Lane and Urban Road, suggesting that size alone doesn't guarantee perceptual granularity in urban contexts.

The overall confusion patterns here suggest that scene understanding benefits from architectural scaling, but can be closely approached with efficient fine-tuning.

Agent Classification

In this task, the model must identify key entities such as vehicles, cyclists, pedestrians, and infrastructure. It yielded the most pronounced confusion across all models, particularly for Qwen2-VL-2B, which shows almost complete failure to distinguish between agents (e.g., mislabeling cyclists as infrastructure).

HazardNet substantially reduces this confusion, especially for vulnerable agents like pedestrians and cyclists, reflecting its improved object grounding and contextual reasoning. Still, moderate FP rates remain, indicating room for refinement in visual representation or attention mechanisms.

Suggested Action

This is the most complex task, requiring the model to recommend appropriate responses such as slowing down carefully, stopping, yielding, or continuing movement. Despite involving five nuanced classes. Despite being the most complex task (with five nuanced classes), HazardNet achieves the least class confusion relative to its model size. Notably, it outperforms GPT4o-mini, correctly differentiating between subtle action cues like "Slow Down" vs. "Must Stop."

GPT4o-mini, while accurate overall, misclassifies a significant number of high-risk actions (e.g., labeling "Must Stop" as "Keep Moving"), a dangerous failure mode in real-world deployment. This supports the thesis that smaller, purpose-trained models can surpass larger generalists in critical decision-making tasks.

4.5 ROC Curve Analysis

As shown in figure 4.4, the ROC curves validate HazardNet's capability to reliably distinguish positive and negative cases across a wide range of tasks. While GPT4o achieves strong performance due to its sheer size, HazardNet achieves comparable or superior AUC in key tasks, such as *Suggested Action* and *Risk Identification*, demonstrating that domain specialization via efficient fine-tuning can rival , and sometimes surpass , general-purpose large models.

These findings reinforce the broader theme of this chapter: that efficiency and accuracy are not mutually exclusive in safety-critical AI. Instead, when paired with task-specific adaptation, small, well-targeted models like HazardNet offer a viable, deployable alternative to their much larger counterparts.

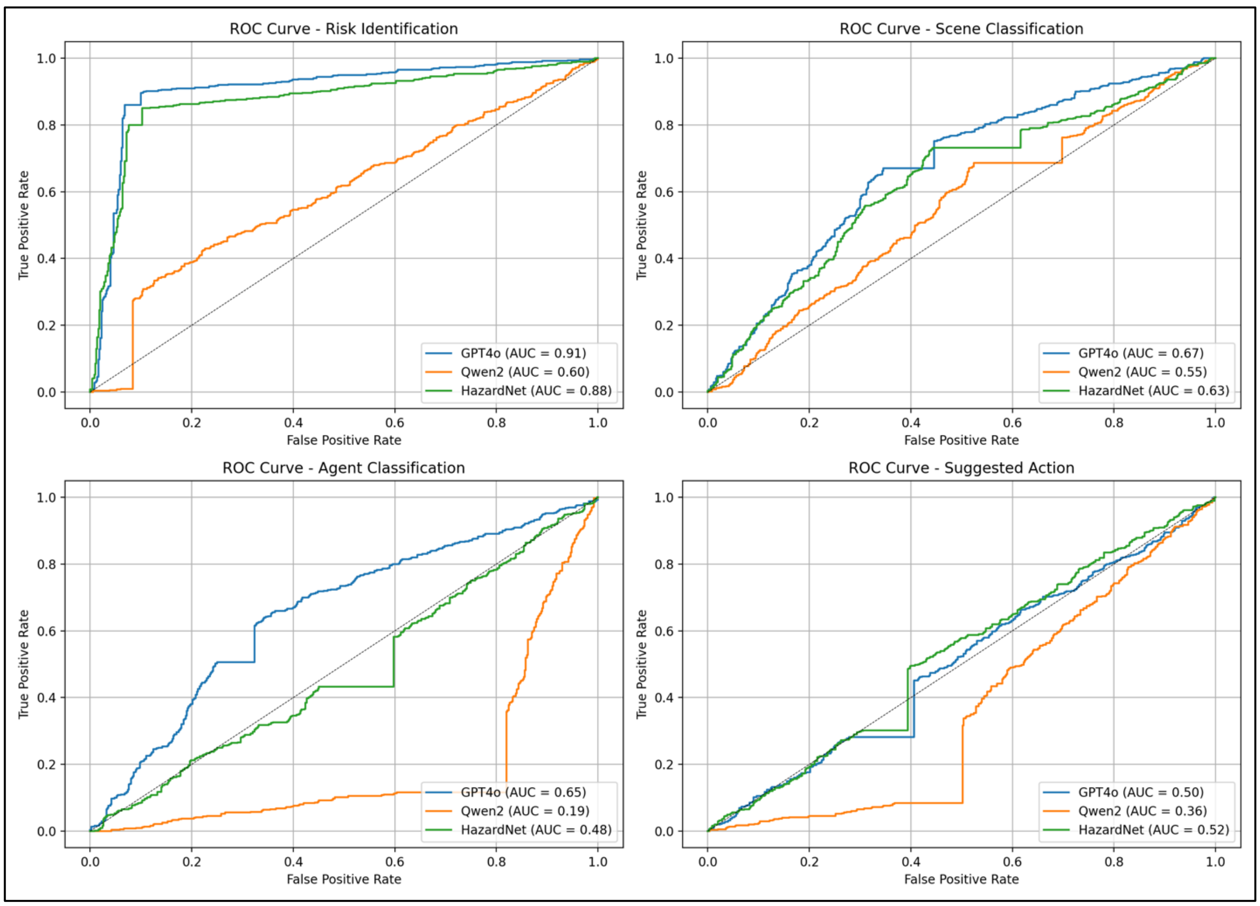


Figure 4.4 ROC curves of evaluated models.

4.6 Qualitative Error Analysis: Understanding Model Failures in Real-World Scenarios

To ensure robust and reliable performance in real-world traffic safety applications, it is critical to examine *how* and *why* HazardNet succeeds or fails in diverse driving scenarios. By conducting a qualitative error analysis, authoraim to:

1. Identify systematic failure modes that could compromise safety in deployment.
2. Understand the model's reasoning gaps, such as over-reliance on certain visual cues (e.g., traffic density) while neglecting others (e.g., contextual traffic signals).
3. Refine HazardNet's decision-making by addressing errors that could lead to unsafe recommendations (e.g., failing to stop at a red light).

Author analyzes six representative traffic scenes (Figure 4.5), spanning low-risk to high-risk conditions, each annotated with ground-truth and predicted outputs across four subtasks: *Scene Type, Agent, Suggested Action,* and *Risk Identification*. These cases were selected to highlight:

1. Common pitfalls (e.g., misinterpreting contextual cues like traffic lights).
2. Overcautious predictions (e.g., flagging non-hazardous congestion as risky).
3. Correct but critical successes (e.g., accurately assessing clear roads).

This analysis not only diagnoses limitations but also informs targeted improvements to enhance HazardNet's safety-critical reasoning. By revealing patterns of misclassification, such as frequent confusion between "Carefully Slow Down" and "Must Stop," or between "Urban Road" and "Narrow Lane", this process identifies where the model struggles to generalize from visual input to correct semantic interpretations. Moreover, these cases uncover areas where the model's confidence does not align with its correctness, suggesting the need for better calibration. Scenes involving pedestrians or cyclists, for instance, show that the model may underperform when objects are partially occluded or when motion needs to be inferred from context, indicating a gap in agent recognition and temporal reasoning.

Furthermore, the presence of consistent errors within specific subtasks allows for focused architectural or training adjustments, such as enhancing multimodal attention mechanisms or expanding the diversity of the training dataset with more edge-case scenarios (e.g., narrow lanes with pedestrian congestion). This kind of analysis also encourages refinement of the label taxonomy, for example, by reconsidering how finely actions should be categorized to reduce ambiguity. Ultimately, qualitative error analysis offers a roadmap for systematically improving HazardNet, not only by fixing individual errors but by evolving the model’s capacity for contextual understanding and safe decision-making.

Case Study (a): Correct Prediction Across All Tasks

This scene features a clear, well-marked urban road with no immediate hazards. HazardNet correctly identifies the scene type, absence of risk, and makes no action recommendation. The model demonstrates strong perceptual grounding and high confidence in low-risk conditions.

Case Study (b): Misclassification of Action

While the model correctly detects the presence of risk and the scene type (narrow lane), it fails to recommend **“Must Stop”**, instead suggesting **“Carefully Slow Down”**. This error likely stems from the model’s limited ability to interpret contextual cues like traffic lights. The red light is present but may not be visually salient enough for the model to infer a mandatory stop, especially amid competing visual elements like cones and signage.

Case Study (c): Incorrect Risk Identification

Despite the congested traffic, this scene poses no imminent danger. However, the model incorrectly flags it as risky. This suggests that HazardNet sometimes over-weights visual density (number of vehicles) without sufficient reasoning about behavioral cues (e.g., spacing, movement), leading to false positives in low-risk but crowded environments.

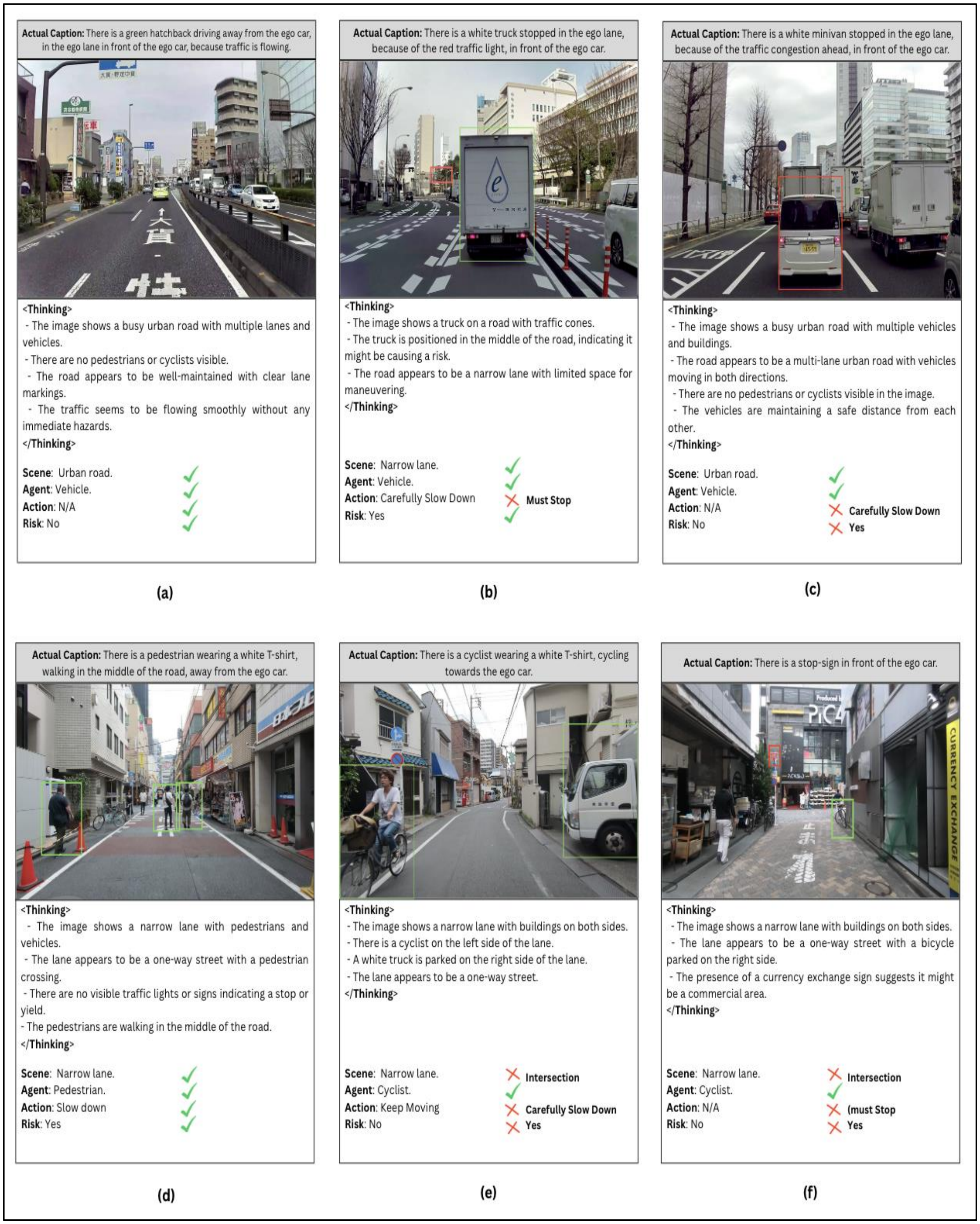

Actual Caption: There is a green hatchback driving away from the ego car, in the ego lane in front of the ego car, because traffic is flowing.
- The image shows a busy urban road with multiple lanes and vehicles.
- There are no pedestrians or cyclists visible.
- The road appears to be well-maintained with clear lane markings.
- The traffic seems to be flowing smoothly without any immediate hazards.
Scene: Urban road.
Agent: Vehicle.
Action: N/A
Risk: No
(a)
Actual Caption: There is a white truck stopped in the ego lane, because of the red traffic light, in front of the ego car.
- The image shows a truck on a road with traffic cones.
- The truck is positioned in the middle of the road, indicating it might be causing a risk.
- The road appears to be a narrow lane with limited space for maneuvering.
Scene: Narrow lane.
Agent: Vehicle.
Action: Carefully Slow Down
Must Stop
Risk: Yes
(b)
Actual Caption: There is a white minivan stopped in the ego lane, because of the traffic congestion ahead, in front of the ego car.
- The image shows a busy urban road with multiple vehicles and buildings.
- The road appears to be a multi-lane urban road with vehicles moving in both directions.
- There are no pedestrians or cyclists visible in the image.
- The vehicles are maintaining a safe distance from each other.
Scene: Urban road.
Agent: Vehicle.
Action: N/A
Carefully Slow Down
Risk: No
Yes
(c)
Actual Caption: There is a pedestrian wearing a white T-shirt, walking in the middle of the road, away from the ego car.
- The image shows a narrow lane with pedestrians and vehicles.
- The lane appears to be a one-way street with a pedestrian crossing.
- There are no visible traffic lights or signs indicating a stop or yield.
- The pedestrians are walking in the middle of the road.
Scene: Narrow lane.
Agent: Pedestrian.
Action: Slow down
Risk: Yes
(d)
Actual Caption: There is a cyclist wearing a white T-shirt, cycling towards the ego car.
- The image shows a narrow lane with buildings on both sides.
- There is a cyclist on the left side of the lane.
- A white truck is parked on the right side of the lane.
- The lane appears to be a one-way street.
Scene: Narrow lane.
Intersection
Agent: Cyclist.
Action: Keep Moving
Carefully Slow Down
Risk: No
Yes
(e)
Actual Caption: There is a stop-sign in front of the ego car.
- The image shows a narrow lane with buildings on both sides.
- The lane appears to be a one-way street with a bicycle parked on the right side.
- The presence of a currency exchange sign suggests it might be a commercial area.
Scene: Narrow lane.
Intersection
Agent: Cyclist.
Action: N/A
(must Stop
Risk: No
Yes
(f)

Figure 4.5 Qualitative error analysis of HazardNet's predictions across six real-world traffic scenes. Each subfigure (a–f) includes the actual caption, the model's predicted class labels (Scene, Agent, Action, Risk), and the correct labels.

Case Study (d): Correct Identification of Pedestrian Risk

The model successfully identifies the presence of pedestrians walking mid-road and classifies the scene as risky. This shows that HazardNet can handle complex pedestrian scenarios in narrow, shared spaces, particularly when the visual cues are prominent and the agent class is clearly exposed.

Case Study (e): Underestimation of Cyclist Risk

Although a cyclist is approaching in a narrow lane, the model incorrectly concludes that no risk is present. This illustrates a potential blind spot in detecting dynamic threats in constrained environments. HazardNet may struggle to infer trajectory or collision potential , critical for interpreting interactions with vulnerable agents like cyclists.

Case Study (f): Misinterpretation of Stop Sign and Risk

In this case, a visible stop sign is present in the background, but it does not apply to the ego vehicle. HazardNet incorrectly predicts **"Must Stop"** and assigns a **risk** label, despite the absence of an immediate hazard. This highlights a broader challenge in *contextual localization* distinguishing whether signage is relevant to the ego agent or applies to intersecting lanes.

Key Takeaways from Qualitative Analysis

Our analysis reveals that HazardNet demonstrates strong performance in static risk detection (e.g., identifying stopped vehicles or pedestrian presence), but faces several interpretative challenges when processing complex traffic scenarios. The model occasionally misinterprets indirect visual cues like traffic lights or signage due to their context-dependent nature, particularly when these elements appear ambiguous or occluded. In scenes with visual clutter or dense traffic, author observe instances of overgeneralization where the system tends to over-penalize risks based on surface features alone. A more significant limitation emerges in temporal reasoning, where the model struggles with trajectory prediction and intent forecasting for dynamic agents like cyclists - what author term "trajectory blindness." Additionally, spatial grounding proves challenging, as evidenced by sign relevance errors where the model misattributes traffic signs to the ego lane. These findings collectively highlight that while HazardNet provides reliable static hazard detection, its performance degrades when processing dynamic or context-sensitive elements. Addressing these limitations through enhanced temporal modeling (potentially via video inputs or motion cues) and improved spatial attention mechanisms represents a critical direction for future work to advance robust safety-critical reasoning in autonomous systems.

4.7 Dataset Impact on Learning

HazardQA consists of image–question–answer triplets, each enriched with an intermediate chain-of-thought rationale that explains the reasoning behind the correct answer. This structure allows the model not only to learn what the correct answer is, but why that answer is appropriate in the context of real-world driving.

HazardQA contains over 7,000 annotated examples spanning a broad spectrum of visual reasoning categories relevant to autonomous driving. Each sample includes:

1. Question: A natural-language query related to the image.
2. Chain of Thought: A short explanatory rationale articulating visual cues and logic.
3. Answer: The correct response based on the image and reasoning.

The inclusion of explicit *chain-of-thought* rationales proved especially valuable during fine-tuning. Rather than relying purely on image-text correlations, HazardNet was able to learn the intermediate reasoning steps that connect visual observations to safety-critical decisions. This contributed to:

1. Improved Generalization: By seeing rationales across diverse categories, the model developed transferable reasoning skills (e.g., linking pedestrian proximity with required action).

2. Safer Decision-Making**:** Tasks such as Suggested Action and Risk Identification benefited from examples where causality and spatial logic were explicitly explained.
3. Interpretable Learning: The chain-of-thought format aligns naturally with instruction tuning and explainable AI goals, supporting future human-in-the-loop validation.

In summary, HazardQA not only provided diverse and realistic visual scenarios but also reinforced multi-step reasoning patterns that enabled HazardNet to perform well in high-stakes tasks requiring interpretability and contextual understanding.

4.8 Architecture–Performance Synergy

The strong performance of HazardNet, particularly in Risk Identification and Suggested Action tasks, can be directly attributed to the architectural strengths of its foundation model, Qwen2-VL-2B. Figure 4.6 illustrates the architecture, which features a modular dual-tower design consisting of a high-capacity Vision Encoder and a QwenLM Decoder, capable of unified visual-textual reasoning. This architecture played a pivotal role in enabling effective adaptation to the HazardQA dataset during fine-tuning.

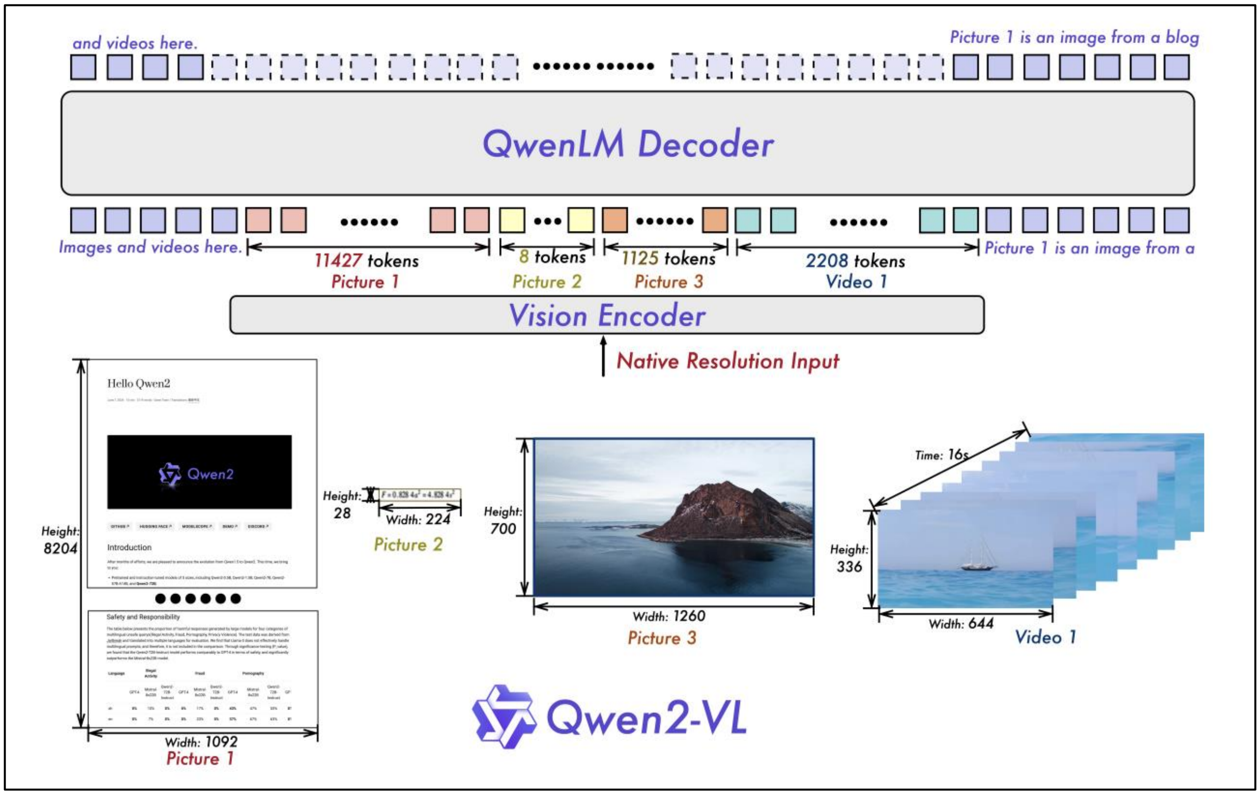


Figure 4.6 Qwen2-VL architecture (Wang et al., 2024).

Native Resolution Input for Fine-Grained Perception

A unique advantage of Qwen2-VL is its ability to accept native-resolution visual inputs, including high-resolution images (e.g., 8204×1092 in Picture 1). This allows the Vision Encoder to preserve critical spatial and temporal cues, such as traffic sign text, pedestrian gestures, or vehicle positioning, that are often lost in aggressive down sampling.

In the context of HazardQA, this was particularly impactful for:

1. Traffic Rules and Compliance: Fine text and signage were preserved (e.g., stop signs, lane indicators).
2. Spatial Relationships and Positioning: The model could resolve narrow lanes, parked vehicles, and pedestrian crossings with higher spatial fidelity.

Token-Efficient Multimodal Fusion

Visual elements are encoded into compact token sequences (e.g., Picture 2 → 8 tokens, Picture 3 → 1125 tokens), enabling efficient alignment with textual tokens in the decoder. This design supports the integration of long chain-of-thought rationales from HazardQA without exceeding token limits, allowing the model to simultaneously attend to visual details and logical reasoning steps.

As a result, HazardNet was able to:

1. Learn causal relationships between observations and actions (e.g., "The light is red → Must Stop").
2. Maintain context continuity when reasoning over multi-element scenes (e.g., dynamic movement near intersections).

Decoder-Based Reasoning with Chain-of-Thought Conditioning

The QwenLM Decoder, pretrained for language generation, was highly compatible with the chain-of-thought format of HazardQA. During fine-tuning, it was exposed to rationales for safety-related decisions, enabling the model to internalize patterns of logic that govern driving behavior.

This alignment enabled:

1. High performance in Risk Identification (F1-score: 84.10%) by modeling stepwise reasoning about threats.
2. Better generalization to novel scenarios, where explicit rationale helped the model extrapolate decision logic beyond training distribution.

4.9 Prompt Analysis

Prompt engineering played a foundational role in both the creation of the HazardQA dataset and the evaluation of HazardNet. Carefully designed instruction formats enabled the model to engage in structured reasoning during training and consistent inference during evaluation. This section analyzes the two core prompts used in the system pipeline and discusses their design choices, rationale, and downstream impacts on performance.

4.9.1 Prompt for HazardQA Dataset Creation

The first prompt (Appendix A.1) was used to generate the HazardQA dataset from visual traffic scenes. This prompt was designed with the goal of eliciting not only correct answers but also interpretable reasoning traces that model how humans process safety-relevant visual information. Key features include:

1. Structured Output Schema: The prompt enforces a JSON format with explicit fields for `question`, `category_topic`, `chain_of_thought`, and `answer`, ensuring consistency and machine-readability across thousands of generated examples.
2. Diverse Reasoning Categories: Questions are distributed across seven high-level reasoning categories, ranging from Object Detection to Scene Prediction, promoting broad cognitive coverage across perception, inference, and context modeling.
3. Explicit Chain-of-Thought Reasoning: By requiring a rationale for each answer, the prompt enforces a format that mirrors human cognitive processes. This was critical for training HazardNet to not only recognize features but understand their implications (e.g., "The light is red → Must Stop").

Impact on Training:

This prompt enabled the generation of a reasoning-centric dataset that trained HazardNet to internalize cause-effect relationships in driving scenes. The inclusion of intermediate rationales improved generalization and interpretability, especially in complex tasks like Suggested Action and Risk Identification.

4.9.2 Prompt for HazardNet Evaluation

The second prompt (Appendix A.2) was used to evaluate HazardNet's end-to-end reasoning and decision-making capabilities. This evaluation prompt mirrors real-world autonomy use cases, requiring the model to synthesize multiple cognitive outputs in a structured XML format. Key design choices include:

1. Modular Reasoning Flow: The prompt breaks down the model’s response into five explicit subcomponents, `Thinking`, `SceneClassification`,

`AgentClassification`, `SuggestedAction`, and `RiskAssessment`, allowing for fine-grained analysis of each cognitive step.

2. Instructional Constraints: The prompt enforces strict output formatting using `<SceneAnalysis>` tags and prohibits extraneous commentary. This constraint ensures consistency across predictions, enabling easy parsing and quantitative evaluation.
3. Conditional Logic: Embedded rules (e.g., "If Risk = No, then Suggested Action = N/A") introduce procedural logic into the reasoning chain, encouraging models to learn conditional dependencies rather than isolated outputs.

Impact on Evaluation:

This prompt not only enforced structural discipline in outputs but also revealed nuanced model behaviors under controlled reasoning instructions. By requiring a coherent Thinking step, the prompt served as an implicit test of HazardNet's internal logic, allowing qualitative analysis of both correct predictions and failure cases (as shown in Section 4.6).

The careful co-design of dataset generation and evaluation prompts created a closed-loop supervision framework where the reasoning logic embedded during training (via chain-of-thoughts) could be assessed consistently at inference time. This tight prompt alignment contributed to HazardNet's strong performance in complex reasoning tasks, and highlights the broader importance of prompt engineering as a mechanism for shaping model cognition in vision-language domains.

4.10 Summary

This chapter presented a comprehensive evaluation of HazardNet's performance across a suite of safety-critical visual question answering (VQA) tasks relevant to autonomous driving. Leveraging parameter-efficient fine-tuning methods (LoRA/QLoRA), HazardNet demonstrated strong task-specific generalization, achieving near-parity with the 20× larger GPT4o-mini while significantly outperforming the base Qwen2-VL-2B model.

Quantitative results across tasks, including Risk Identification, Scene Classification, Agent Classification, and Suggested Action, highlighted HazardNet's ability to balance precision, recall, and efficiency, delivering competitive F1-scores and ROC-AUC values. Notably, in high-risk decision contexts like Suggested Action, HazardNet surpassed GPT4o, validating the effectiveness of targeted specialization over raw scale.

Confusion matrix analysis revealed meaningful reductions in class confusion, particularly in nuanced multi-class tasks, while ROC curves affirmed HazardNet's strong discriminative capacity across thresholds. Together, these findings demonstrate that compact, optimized models can meet the stringent reliability requirements of autonomous systems without sacrificing performance or interpretability.

Chapter Five: Conclusion and Future Work

5.1 Conclusion

This thesis addressed the critical challenge of creating efficient, scalable, and reliable AI systems capable of real-time safety-critical event detection for autonomous driving. By leveraging advancements in multimodal language models (MLLMs) and parameter-efficient fine-tuning methods, author introduced HazardNet, a compact vision-language model designed specifically for traffic safety, and developed HazardQA, a novel Vision Question Answering (VQA) dataset tailored for reasoning in high-risk driving scenarios.

The research yielded several key findings. First, HazardNet achieved an impressive 84.10% F1-score in Risk Identification tasks, closing 97% of the performance gap with the much larger GPT4o-mini model, which is twenty times its size. This result validates the effectiveness of parameter-efficient techniques like Low-Rank Adaptation (LoRA) and Quantized LoRA (QLoRA) in allowing smaller models to approach the performance of their larger counterparts while retaining computational efficiency. Second, the HazardQA dataset, comprising 7,125 question-answer pairs generated by GPT-4o and validated by human annotators, offered a robust framework for training models to reason about traffic hazards, context, and dynamic elements in driving environments. Third, the parameter-efficient design of HazardNet, requiring updates to just 0.16% of parameters and using only 12.1 GB of VRAM, demonstrated the model's viability for deployment on edge devices, with inference speeds suitable for real-time applications.

This thesis successfully addressed the four research questions outlined in Chapter 1. First, through extensive evaluation, it demonstrated that HazardNet performs competitively with much larger models like GPT-4o-mini, confirming the effectiveness of fine-tuned small-scale MLLMs in safety-critical event detection. Second, it presented a detailed methodology for constructing HazardQA, including data generation, annotation, and integration into training pipelines effectively answering how domain-specific VQA datasets can support hazard reasoning. Third, the thesis tackled the challenges of edge deployment by optimizing model size, memory usage, and inference latency, proving the feasibility of real-time performance on low-resource hardware. Finally, it showed how HazardNet's ability to deliver accurate and timely hazard detection can support proactive traffic safety management, fulfilling its intended role in enhancing real-world autonomous systems.

Collectively, these outcomes confirm that HazardNet effectively meets the goals outlined in Chapter 1. The model strikes a balance between accuracy, interpretability, and efficiency, making it a strong candidate for safety-critical applications in autonomous systems.

5.2 Contributions and Possible Applications

This work introduced several notable contributions. First, HazardNet represents a lightweight, fine-tuned MLLM derived from Qwen2-VL-2B, optimized for hazard detection through LoRA and QLoRA. Its architecture supports efficient deployment on resource-constrained edge devices while retaining strong multimodal reasoning capabilities. Second, the HazardQA dataset stands as the first VQA resource focused exclusively on safety-critical traffic scenarios. By offering structured QA pairs and chain-of-thought explanations, it bridges the gap between visual perception and actionable insights. Third, this research demonstrated a methodological innovation in applying parameter-efficient fine-tuning to adapt large models to niche domains. This approach significantly reduces training time and hardware requirements without sacrificing performance.

The potential applications of this work are broad and impactful. In the context of autonomous vehicles, HazardNet can enhance real-time hazard detection and support better decision-making in complex, high-risk scenarios, such as navigating intersections or driving in adverse weather conditions. In Advanced Driver-Assistance Systems (ADAS), the model offers contextual reasoning capabilities that can improve collision warnings, pedestrian detection, and risk assessment. Beyond vehicles, edge computing environments, such as traffic cameras and IoT-enabled infrastructure in smart cities, can benefit from deploying HazardNet to monitor traffic hazards efficiently. Additionally, HazardQA can serve as a valuable benchmarking tool for evaluating VQA models designed for safety-centric applications, encouraging further research and community-driven development.

5.3 Future Work

Several directions can be pursued to extend this research. One promising avenue is the creation of a video-based version of HazardQA to support temporal reasoning, addressing the limitations of static frame-based models when it comes to interpreting motion and intent. Another priority is enhancing the geographic and scenario diversity of the dataset by incorporating data from underrepresented regions and edge cases, such as extreme weather conditions or sensor malfunctions.

Improving annotation quality is also key. Incorporating a human-in-the-loop approach to supplement synthetic QA pairs with human-generated content can increase linguistic diversity and reduce biases introduced by large language models like GPT-4o. On the system integration side, real-time integration of HazardNet with sensor data streams (e.g., LiDAR, radar) and vehicle-to-everything (V2X) communication could enable more holistic and context-aware hazard prediction. From a model design perspective, future work may explore architectural enhancements, such as combining HazardNet with reinforcement learning to improve its ability to recommend actions, an area that showed weaker performance in current evaluations. Finally, to ensure broader applicability, further testing on a range of edge hardware platforms, such as the Jetson Nano or Raspberry Pi, will help optimize throughput and latency for diverse real-world deployment scenarios.

5.4 Closing Remarks

The development of HazardNet and HazardQA marks a significant advancement in democratizing AI-powered traffic safety solutions. This work challenges the prevailing assumption that only large models can deliver high performance, showing instead that with domain-specific optimization and efficient fine-tuning, smaller models can achieve comparable results. As autonomous systems become increasingly integrated into modern transportation, solutions like HazardNet will be vital in reducing traffic-related injuries and fatalities, minimizing economic losses, and supporting the transition to safer, smarter mobility systems.

Looking forward, the challenge will be to ensure that technological advancements remain aligned with real-world deployability. Continued research and development must prioritize practicality, accessibility, and robustness, ensuring that AI not only advances in capability but also becomes a dependable partner in achieving safer roads worldwide.

References

Abibullaev, B., Keutayeva, A., & Zollanvari, A. (2023). Deep Learning in EEG-Based BCIs: A Comprehensive Review of Transformer Models, Advantages, Challenges, and Applications. *IEEE Access*, *11*, 127271–127301. https://doi.org/10.1109/ACCESS.2023.3329678

Abu Tami, M., Ashqar, H. I., Elhenawy, M., Glaser, S., & Rakotonirainy, A. (2024). Using Multimodal Large Language Models (MLLMs) for Automated Detection of Traffic Safety-Critical Events. *Vehicles*, *6*(3), 1571–1590. https://doi.org/10.3390/vehicles6030074

Adewopo, V., Elsayed, N., Elsayed, Z., Ozer, M., Zekios, C. L., Abdelgawad, A., & Bayoumi, M. (2024). Big Data and Deep Learning in Smart Cities: A Comprehensive Dataset for AI-Driven Traffic Accident Detection and Computer Vision Systems. *SoutheastCon 2024*, 675–680. https://doi.org/10.1109/SoutheastCon52093.2024.10500288

Bai, Z., Wang, P., Xiao, T., He, T., Han, Z., Zhang, Z., & Shou, M. Z. (2024). *Hallucination of Multimodal Large Language Models: A Survey*.

Bansal, T., Jha, R., & McCallum, A. (2020). Learning to Few-Shot Learn Across Diverse Natural Language Classification Tasks. *Proceedings of the 28th International Conference on Computational Linguistics*, 5108–5123. https://doi.org/10.18653/v1/2020.coling-main.448

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., … Amodei, D. (2020). *Language Models are Few-Shot Learners*.

Caesar, H. V. B. A. H. L. S. V. V. E. L. Q. X. A. K. Y. P. G. B. and O. Beijbom. (2020). nuscenes: A multimodal dataset for autonomous driving. *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11621–11631.

Cao, F., Chen, S., Zhong, J., & Gao, Y. (2023). Traffic Condition Classification Model Based on Traffic-Net. *Computational Intelligence and Neuroscience*, *2023*(1). https://doi.org/10.1155/2023/7812276

Cao, Y., Xiao, C., Cyr, B., Zhou, Y., Park, W., Rampazzi, S., Chen, Q. A., Fu, K., & Mao, Z. M. (2019). Adversarial Sensor Attack on LiDAR-based Perception in Autonomous Driving. *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*, 2267–2281. https://doi.org/10.1145/3319535.3339815

Chen, D. and K. P. (2022). Learning From All Vehicles. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, *6*, 17222–17231.

Chen, L., Sinavski, O., Hünermann, J., Karnsund, A., Willmott, A. J., Birch, D., Maund, D., & Shotton, J. (2024). Driving with LLMs: Fusing Object-Level Vector Modality for Explainable Autonomous Driving. *2024 IEEE International Conference on Robotics and Automation (ICRA)*, 14093–14100. https://doi.org/10.1109/ICRA57147.2024.10611018

Chen, Y., Tonkens, S., & Pavone, M. (2023). *Categorical Traffic Transformer: Interpretable and Diverse Behavior Prediction with Tokenized Latent*.

Chen, Z., Wu, J., Wang, W., Su, W., Chen, G., Xing, S., Zhong, M., Zhang, Q., Zhu, X., Lu, L., Li, B., Luo, P., Lu, T., Qiao, Y., & Dai, J. (2023). *InternVL: Scaling up Vision Foundation Models and Aligning for Generic Visual-Linguistic Tasks*.

Cordts, M. and O. M. and R. S. and R. T. and E. M. and B. R. and F. U. and R. S. and S. B. (2016). The Cityscapes Dataset for Semantic Urban Scene Understanding. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)}*, *6*.

Cui, C. and M. Y. and C. X. and Y. W. and W. Z. (2024). Drive As You Speak: Enabling Human-Like Interaction With Large Language Models in Autonomous Vehicles. *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) Workshops*, 902–909.

Cui, C., Ma, Y., Cao, X., Ye, W., Zhou, Y., Liang, K., Chen, J., Lu, J., Yang, Z., Liao, K.-D., Gao, T., Li, E., Tang, K., Cao, Z., Zhou, T., Liu, A., Yan, X., Mei, S., Cao, J., … Zheng, C. (2024). A Survey on Multimodal Large Language Models for Autonomous Driving. *2024 IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW)*, 958–979. https://doi.org/10.1109/WACVW60836.2024.00106

Dettmers, T. and P. A. and H. A. and Z. L. (2024). Qlora: Efficient finetuning of quantized llms. *Advances in Neural Information Processing Systems*, *35*(dettmers2024qlora).

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2020). *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale*.

Eykholt, K. and E. I. and F. E. and L. B. and R. A. and X. C. and P. A. and K. T. and S. D. (2018). Robust Physical-World Attacks on Deep Learning Visual Classification. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Fu, L., Yang, B., Kuang, Z., Song, J., Li, Y., Zhu, L., Luo, Q., Wang, X., Lu, H., Huang, M., Li, Z., Tang, G., Shan, B., Lin, C., Liu, Q., Wu, B., Feng, H., Liu, H., Huang, C., … Bai, X. (2024). *OCRBench v2: An Improved Benchmark for Evaluating Large Multimodal Models on Visual Text Localization and Reasoning*.

Geiger, A., Lenz, P., Stiller, C., & Urtasun, R. (2013). Vision meets robotics: The KITTI dataset. *The International Journal of Robotics Research*, *32*(11), 1231–1237. https://doi.org/10.1177/0278364913491297

Gemini Team, Anil, R., Borgeaud, S., Alayrac, J.-B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A. M., Hauth, A., Millican, K., Silver, D., Johnson, M., Antonoglou, I., Schrittwieser, J., Glaese, A., Chen, J., Pitler, E., Lillicrap, T., Lazaridou, A., … Vinyals, O. (2023). *Gemini: A Family of Highly Capable Multimodal Models*.

Guan, J., Ding, T., Cao, L., Pan, L., Wang, C., & Zheng, X. (2024). *Probing the Robustness of Vision-Language Pretrained Models: A Multimodal Adversarial Attack Approach*.

Hankey, J. M. , P. M. A. and M. J. A. (2016). Description of the SHRP 2 naturalistic database and the crash, near-crash, and baseline data sets. *Virginia Tech Transportation Institute*.

Hanselmann, N., Renz, K., Chitta, K., Bhattacharyya, A., & Geiger, A. (2022). *KING: Generating Safety-Critical Driving Scenarios for Robust Imitation via Kinematics Gradients* (pp. 335–352). https://doi.org/10.1007/978-3-031-19839-7_20

Howarth, J. (2024, April 4). *Number of Parameters in GPT-4 (Latest Data)*.
Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., & Chen, W. (2021). *LoRA: Low-Rank Adaptation of Large Language Models*.
Hu, S., Tu, Y., Han, X., He, C., Cui, G., Long, X., Zheng, Z., Fang, Y., Huang, Y., Zhao, W., Zhang, X., Thai, Z. L., Zhang, K., Wang, C., Yao, Y., Zhao, C., Zhou, J., Cai, J., Zhai, Z., … Sun, M. (2024). *MiniCPM: Unveiling the Potential of Small Language Models with Scalable Training Strategies*.
Hu, Y., Ye, D., Kang, J., Wu, M., & Yu, R. (2024). A Cloud-Edge Collaborative Architecture for Multimodal LLMs-Based Advanced Driver Assistance Systems in IoT Networks. *IEEE Internet of Things Journal*, 1–1. https://doi.org/10.1109/JIOT.2024.3509628
Huang, Z., Sheng, Z., Qu, Y., You, J., & Chen, S. (2024). *VLM-RL: A Unified Vision Language Models and Reinforcement Learning Framework for Safe Autonomous Driving*.
Jaradat, S., Nayak, R., & Elhenawy, M. (2024). *Advanced Traffic Safety Analysis: Leveraging Deep Learning and Large Language Models for Near-Crash Detection in Crowdsourced Videos* (pp. 495–513). https://doi.org/10.1007/978-3-031-73125-9_32
Jarrahi, M. H. (2018). Artificial intelligence and the future of work: Human-AI symbiosis in organizational decision making. *Business Horizons*, *61*(4), 577–586. https://doi.org/10.1016/j.bushor.2018.03.007
Jin, B., Liu, X., Zheng, Y., Li, P., Zhao, H., Zhang, T., Zheng, Y., Zhou, G., & Liu, J. (2023). ADAPT: Action-aware Driving Caption Transformer. *2023 IEEE International Conference on Robotics and Automation (ICRA)*, 7554–7561. https://doi.org/10.1109/ICRA48891.2023.10160326
Jin, Y., Yang, R., Yi, Z., Shen, X., Peng, H., Liu, X., Qin, J., Li, J., Xie, J., Gao, P., Zhou, G., & Gong, J. (2024). SurrealDriver: Designing LLM-powered Generative Driver Agent Framework based on Human Drivers' Driving-thinking Data. *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 966–971. https://doi.org/10.1109/IROS58592.2024.10802229
Kalamkar, S., & A., G. M. (2023). Multimodal image fusion: A systematic review. *Decision Analytics Journal*, *9*, 100327. https://doi.org/10.1016/j.dajour.2023.100327
Khan, A., Rauf, Z., Sohail, A., Khan, A. R., Asif, H., Asif, A., & Farooq, U. (2023). A survey of the vision transformers and their CNN-transformer based variants. *Artificial Intelligence Review*, *56*(S3), 2917–2970. https://doi.org/10.1007/s10462-023-10595-0
Komasi, H., Nemati, A., Hashemkhani Zolfani, S., & Mehtari Taheri, H. (2024). Road safety evaluation in inner-city roads and suburban roads based on a novel-hybrid MCDM model. *Ain Shams Engineering Journal*, *15*(8). https://doi.org/10.1016/J.ASEJ.2024.102796
Li, H., Chen, J., Wei, Z., Huang, S., Hui, T., Gao, J., Wei, X., & Liu, S. (2025). *LLaVA-ST: A Multimodal Large Language Model for Fine-Grained Spatial-Temporal Understanding*.
Lin, Y., Luo, L., Chen, Y., Zhang, X., Wang, Z., Yang, W., Tong, M., & Yu, R. (2024). *ST-Align: A Multimodal Foundation Model for Image-Gene Alignment in Spatial Transcriptomics*.
Liu, H., Li, C., Wu, Q., & Lee, Y. J. (2023). *Visual Instruction Tuning*.
Liu, L., Cheng, Y., Deng, Z., Wang, S., Chen, D., Hu, X., Liò, P., Schönlieb, C.-B., & Aviles-Rivero, A. (2024). TrafficMOT: A Challenging Dataset for Multi-Object

Tracking in Complex Traffic Scenarios. *Proceedings of the 32nd ACM International Conference on Multimedia*, 1265–1273. https://doi.org/10.1145/3664647.3681153

Liu, S., Liu, L., Tang, J., Yu, B., Wang, Y., & Shi, W. (2019). Edge Computing for Autonomous Driving: Opportunities and Challenges. *Proceedings of the IEEE*, *107*(8), 1697–1716. https://doi.org/10.1109/JPROC.2019.2915983

Liu, S., Pu, W., Xu, C., Huang, Z., Li, Q., Wang, H., Lin, C., & Shen, C. (2024). *A Comprehensive Survey of Multimodal Large Language Models: Concept, Application and Safety*. https://doi.org/10.21203/rs.3.rs-5270567/v1

Mahesh, B. (2020). Machine Learning Algorithms - A Review. *International Journal of Science and Research (IJSR)*, *9*(1), 381–386. https://doi.org/10.21275/ART20203995

Mahmud, D., Hajmohamed, H., Almentheri, S., Alqaydi, S., Aldhaheri, L., Khalil, R. A., & Saeed, N. (2025). Integrating LLMs With ITS: Recent Advances, Potentials, Challenges, and Future Directions. *IEEE Transactions on Intelligent Transportation Systems*, 1–36. https://doi.org/10.1109/TITS.2025.3528116

Malla, S. and C. C. and D. I. and C. J. H. and L. J. (2023). DRAMA: Joint Risk Localization and Captioning in Driving. *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, *1*, 1043–1052.

Mao, J., Qian, Y., Ye, J., Zhao, H., & Wang, Y. (2023). *GPT-Driver: Learning to Drive with GPT*.

Mathew, M., Karatzas, D., & Jawahar, C. V. (2020). *DocVQA: A Dataset for VQA on Document Images*.

Montiel-Marín, S., Gómez-Huélamo, C., de la Peña, J., Antunes, M., López-Guillén, E., & Bergasa, L. M. (2023). *Towards LiDAR and RADAR Fusion for Object Detection and Multi-object Tracking in CARLA Simulator* (pp. 552–563). https://doi.org/10.1007/978-3-031-21062-4_45

Nam, W., & Jang, B. (2024). A survey on multimodal bidirectional machine learning translation of image and natural language processing. *Expert Systems with Applications*, *235*, 121168. https://doi.org/10.1016/j.eswa.2023.121168

Nippani, A. , L. D. , J. H. , K. H. and Z. H. (2024). Graph neural networks for road safety modeling: datasets and evaluations for accident analysis. *Advances in Neural Information Processing Systems*.

OpenAI, :, Hurst, A., Lerer, A., Goucher, A. P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A., Welihinda, A., Hayes, A., Radford, A., Mądry, A., Baker-Whitcomb, A., Beutel, A., Borzunov, A., Carney, A., Chow, A., Kirillov, A., … Malkov, Y. (2024). *GPT-4o System Card*.

Patil, R., & Gudivada, V. (2024). A Review of Current Trends, Techniques, and Challenges in Large Language Models (LLMs). *Applied Sciences*, *14*(5), 2074. https://doi.org/10.3390/app14052074

Pérez-Castán, J. A., Pérez Sanz, L., Fernández-Castellano, M., Radišić, T., Samardžić, K., & Tukarić, I. (2022). Learning Assurance Analysis for Further Certification Process of Machine Learning Techniques: Case-Study Air Traffic Conflict Detection Predictor. *Sensors*, *22*(19), 7680. https://doi.org/10.3390/s22197680

Qin, H., Ma, X., Zheng, X., Li, X., Zhang, Y., Liu, S., Luo, J., Liu, X., & Magno, M. (2024). *Accurate LoRA-Finetuning Quantization of LLMs via Information Retention*.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). *Learning Transferable Visual Models From Natural Language Supervision*.

Razi, A., Chen, X., Li, H., Wang, H., Russo, B., Chen, Y., & Yu, H. (2022). *Deep Learning Serves Traffic Safety Analysis: A Forward-looking Review*.

Ren, H., Huang, T., & Yan, H. (2021). Adversarial examples: attacks and defenses in the physical world. *International Journal of Machine Learning and Cybernetics*, *12*(11), 3325–3336. https://doi.org/10.1007/s13042-020-01242-z

Sha, H., Mu, Y., Jiang, Y., Chen, L., Xu, C., Luo, P., Li, S. E., Tomizuka, M., Zhan, W., & Ding, M. (2023). *LanguageMPC: Large Language Models as Decision Makers for Autonomous Driving*.

Shi, L., Jiang, B., Zeng, T., & Guo, F. (2024a). *ScVLM: Enhancing Vision-Language Model for Safety-Critical Event Understanding*.

Shi, L., Jiang, B., Zeng, T., & Guo, F. (2024b). *ScVLM: Enhancing Vision-Language Model for Safety-Critical Event Understanding*.

Sima, C., Renz, K., Chitta, K., Chen, L., Zhang, H., Xie, C., Beißwenger, J., Luo, P., Geiger, A., & Li, H. (2025). *DriveLM: Driving with Graph Visual Question Answering* (pp. 256–274). https://doi.org/10.1007/978-3-031-72943-0_15

Sohail, A., Cheema, M. A., Ali, M. E., Toosi, A. N., & Rakha, H. A. (2023). Data-driven approaches for road safety: A comprehensive systematic literature review. *Safety Science*, *158*, 105949. https://doi.org/10.1016/J.SSCI.2022.105949

Sun, P. and K. H. and D. X. and C. A. and P. V. and T. P. and G. J. and Z. Y. and C. Y. and C. B. and V. V. and H. W. and N. J. and Z. H. and T. (2020). Scalability in Perception for Autonomous Driving: Waymo Open Dataset. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, *6*.

Suo, S. and R. S. and C. S. and U. R. (2021). TrafficSim: Learning To Simulate Realistic Multi-Agent Behaviors. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 10400–10409.

Tang, J., Liu, Q., Ye, Y., Lu, J., Wei, S., Lin, C., Li, W., Mahmood, M. F. F. Bin, Feng, H., Zhao, Z., Wang, Y., Liu, Y., Liu, H., Bai, X., & Huang, C. (2024). *MTVQA: Benchmarking Multilingual Text-Centric Visual Question Answering*.

Uparkar, O., Bharti, J., Pateriya, R. K., Gupta, R. K., & Sharma, A. (2023). Vision Transformer Outperforms Deep Convolutional Neural Network-based Model in Classifying X-ray Images. *Procedia Computer Science*, *218*, 2338–2349. https://doi.org/10.1016/J.PROCS.2023.01.209

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). *Attention Is All You Need*.

Wang, L., Ren, Y., Jiang, H., Cai, P., Fu, D., Wang, T., Cui, Z., Yu, H., Wang, X., Zhou, H., Huang, H., & Wang, Y. (2024). AccidentGPT: A V2X Environmental Perception Multi-modal Large Model for Accident Analysis and Prevention. *2024 IEEE Intelligent Vehicles Symposium (IV)*, 472–477. https://doi.org/10.1109/IV55156.2024.10588374

Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., Chen, K., Liu, X., Wang, J., Ge, W., Fan, Y., Dang, K., Du, M., Ren, X., Men, R., Liu, D., Zhou, C., Zhou, J., & Lin, J. (2024). *Qwen2-VL: Enhancing Vision-Language Model's Perception of the World at Any Resolution*.

Wang, W., Xie, J., Hu, C., Zou, H., Fan, J., Tong, W., Wen, Y., Wu, S., Deng, H., Li, Z., Tian, H., Lu, L., Zhu, X., Wang, X., Qiao, Y., & Dai, J. (2023). *DriveMLM: Aligning Multi-Modal Large Language Models with Behavioral Planning States for Autonomous Driving*.

World Health Organizatio. (2023, December 13). *Road traffic Injuries*. https://www.who.int/news-room/fact-sheets/detail/road-traffic-injuries

Xiao, Y., Liu, Y., Luan, K., Cheng, Y., Chen, X., & Lu, H. (2023). Deep LiDAR-Radar-Visual Fusion for Object Detection in Urban Environments. *Remote Sensing*, *15*(18), 4433. https://doi.org/10.3390/rs15184433

Xu, Y., Hu, H., Huang, C., Nan, Y., Liu, Y., Wang, K., Liu, Z., & Lian, S. (2025). TAD: A Large-Scale Benchmark for Traffic Accidents Detection From Video Surveillance. *IEEE Access*, *13*, 2018–2033. https://doi.org/10.1109/ACCESS.2024.3522384

Xu, Y., Hu, Y., Zhang, Z., Meyer, G. P., Mustikovela, S. K., Srinivasa, S., Wolff, E. M., & Huang, X. (2024). *VLM-AD: End-to-End Autonomous Driving through Vision-Language Model Supervision*.

Yan, X., Zhang, H., Cai, Y., Guo, J., Qiu, W., Gao, B., Zhou, K., Zhao, Y., Jin, H., Gao, J., Li, Z., Jiang, L., Zhang, W., Zhang, H., Dai, D., & Liu, B. (2024). *Forging Vision Foundation Models for Autonomous Driving: Challenges, Methodologies, and Opportunities*.

Yang, Z., Jia, X., Li, H., & Yan, J. (2023). *LLM4Drive: A Survey of Large Language Models for Autonomous Driving*.

Zhang, D., Yu, Y., Dong, J., Li, C., Su, D., Chu, C., & Yu, D. (2024). MM-LLMs: Recent Advances in MultiModal Large Language Models. *Findings of the Association for Computational Linguistics ACL 2024*, 12401–12430. https://doi.org/10.18653/v1/2024.findings-acl.738

Zhang, R., Wang, B., Zhang, J., Bian, Z., Feng, C., & Ozbay, K. (2025a). *When language and vision meet road safety: leveraging multimodal large language models for video-based traffic accident analysis*.

Zhang, R., Wang, B., Zhang, J., Bian, Z., Feng, C., & Ozbay, K. (2025b). *When language and vision meet road safety: leveraging multimodal large language models for video-based traffic accident analysis*.

Zhang, S., Fu, D., Liang, W., Zhang, Z., Yu, B., Cai, P., & Yao, B. (2024). TrafficGPT: Viewing, processing and interacting with traffic foundation models. *Transport Policy*, *150*, 95–105. https://doi.org/10.1016/j.tranpol.2024.03.006

Zhang, Y., Zhang, K., Li, B., Pu, F., Setiadharma, C. A., Yang, J., & Liu, Z. (2024). *WorldQA: Multimodal World Knowledge in Videos through Long-Chain Reasoning*.

Zhou, X., Liu, M., Yurtsever, E., Zagar, B. L., Zimmer, W., Cao, H., & Knoll, A. C. (2024a). Vision Language Models in Autonomous Driving: A Survey and Outlook. *IEEE Transactions on Intelligent Vehicles*, 1–20. https://doi.org/10.1109/TIV.2024.3402136

Zhou, X., Liu, M., Yurtsever, E., Zagar, B. L., Zimmer, W., Cao, H., & Knoll, A. C. (2024b). Vision Language Models in Autonomous Driving: A Survey and Outlook. *IEEE Transactions on Intelligent Vehicles*, 1–20. https://doi.org/10.1109/TIV.2024.3402136

Appendices

HazardQA Dataset Generation Prompt

```
You are an advanced AI model that analyzes images from a traffic ego
car’s perspective. You will receive an image showing a traffic scene.
Please do the following:

1. **Observe the image carefully** and note the key elements
(vehicles, pedestrians, road markings, signs, etc.).
2. **Generate exactly five (5) Q&A pairs** about the scene.
3. For each Q&A pair:
   - Provide a **Question** that a person might ask about the scene.
   - Identify and include the **Category Topic** for the question,
based on the following predefined list:
     - Object Detection and Recognition
     - Spatial Relationships and Positioning
     - Traffic Rules and Compliance
     - Dynamic Elements and Movement
     - Scene Context and Prediction
     - Weather and Visibility Conditions
     - Road Infrastructure and Features
   - Provide a **Chain-of-Thought** (reasoning steps or thought
process leading to the answer).
   - Provide an **Answer** that directly addresses the question.

4. **Format your output** in valid JSON, structured as follows:

{
  "qa_pairs": [
    {
      "category_topic": "Object Detection and Recognition",
      "question": "Question 1",
      "chain_of_thought": "Reasoning steps.",
      "answer": "final answer."
    },
    {
      "category_topic": "Spatial Relationships and Positioning",
      "question": "Question 2",
      "chain_of_thought": "...",
      "answer": "..."
    },
    ...
  ]
}

Make sure each "chain_of_thought" is clear, and each "answer" is
accurate. Do not include extra commentary beyond these fields.
```

HazardNet Evaluation Prompt

```
You are an advanced AI assistant specialized in analyzing driving
scenes from images. Your task is to process the provided image and
extract detailed information about the scene. Please perform the
following steps:

1. **Thinking**: Include concise reasoning process (chain of thought)
to show how you arrived at your conclusions.

2. **Scene Classification**: Determine the overall type of the scene.
Output one of: [Narrow lane, Intersection, Urban road].

3. **Agent Classification**: Identify agent present in the scene that
might cause the risk. Output one of [N/A, Infrastructure, Pedestrian,
Vehicle, Cyclist].

4. **Suggested Action**: Based on the analysis, recommend the most
appropriate action for the ego vehicle to take, when Risk Assessment
is No, then the Suggested Action is N/A. Output one of [N/A, Yield,
Carefully Slow Down, Must Stop, Keep Moving].

5. **Risk Assessment**: Evaluate whether there is a potential risk in
the current scene that requires immediate attention. Output one of
[Yes, No]

**Response Instructions:**

Start the response **directly with the `<SceneAnalysis>` tag**. Do
not add any explanations, headers, or labels like "XML" or "Output:".
Do not add any descriptive text or explanations.

**Example of the desired response:**
<SceneAnalysis>
    <Thinking></Thinking>
    <SceneClassification></SceneClassification>
    <AgentClassification></AgentClassification>
    <SuggestedAction></SuggestedAction>
    <IsThereRisk></IsThereRisk>
</SceneAnalysis>
```

تعزيز سلامة المرور باستخدام نموذج لغوي كبير متعدد الوسائط للكشف الفوري عن المخاطر.

محمد ياسر أحمد أبو طامع.

د. محمد الحناوي.

د. حذيفة الأشقر.

د. أحمد حساسنة.

ملخص

لا تزال السلامة المرورية تمثل قضية حرجة على مستوى العالم، حيث غالبًا ما تفشل أنظمة الكشف التقليدية في البيئات الواقعية المعقدة بسبب محدودية القدرة على التعميم ومتطلبات الحوسبة العالية. تقدم هذه الرسالة نموذج HazardNet، وهو نموذج لغوي كبير متعدد الوسائط وخفيف الوزن ومتوافق مع الأجهزة الطرفية، تم تحسينه انطلاقًا من نموذج Qwen2-VL-2B باستخدام تقنيات فعالة من حيث عدد المعلمات مثل LoRA و QLoRA. صُمم نموذج HazardNet لاكتشاف المخاطر في الوقت الحقيقي بناءً على السياق، من خلال دمج المدخلات البصرية والنصية، مما يجعله مناسبًا للنشر على الأجهزة منخفضة الموارد مثل الأنظمة داخل المركبات بدون وحدات معالجة رسوميات.(GPU)

لدعم هذا النموذج، تقدم الدراسة مجموعة بيانات جديدة تسمى HazardQA مخصصة لمهام الإجابة على الأسئلة البصرية(VQA) ، تم اشتقاقها من سيناريوهات القيادة الموجودة في مجموعة بيانات DRAMA. تتضمن HazardQA أكثر من 7,000 زوج من الأسئلة والأجوبة المشروحة، مزودة بسلاسل منطقية وتسميات خاصة بالسلامة، وتغطي مجموعة واسعة من المخاطر وسياقات المرور.

تُظهر التجارب أن HazardNet يُحقق أداءً تنافسيًا في المهام الحيوية للسلامة مثل فهم المشهد، وتحديد المخاطر، وتقديم التوصيات للإجراءات، منافسًا نماذج أكبر مثل GPT-4o مع

الحفاظ على متطلبات حوسبة منخفضة. تُبرز هذه الدراسة كيف يمكن للنماذج اللغوية متعددة الوسائط والخفيفة الوزن، عند تكييفها باستخدام بيانات مخصصة للمجال وتقنيات تحسين فعالة، أن توفّر قابلية عالية للتفسير وقابلية للتوسع في أنظمة السلامة المرورية. تشمل المساهمات الرئيسية تطوير نموذج HazardNet، والإصدار مفتوح المصدر لمجموعة بياناتHazardQA ، وعرض استراتيجيات النشر الواقعي لاكتشاف المخاطر باستخدام الذكاء الاصطناعي.

الكلمات المفتاحية: سلامة المرور، نماذج اللغة متعددة الوسائط، الإجابة على الأسئلة البصرية.