Benchmarking Vision-Language Models for Traffic Scene Understanding in South Asian Traffic Environments

By

Naznin Sultana Choity
20301275
Samiha Takmim
21301222
Abida Rahman
21201645
Abdullah Al Shaid
20201103
Md.Tanzim Hossain
21201131

A thesis submitted to the Department of Computer Science and Engineering
in partial fulfillment of the requirements for the degree of
B.Sc. in Computer Science and Engineering

Department of Computer Science and Engineering
BRAC University
December 2025.

Declaration

It is hereby declared that

1. The thesis submitted is my/our own original work while completing degree at Brac University.
2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.
3. The thesis does not contain material which has been accepted, or submitted, for any other degree or diploma at a university or other institution.
4. We have acknowledged all main sources of help.

Student's Full Name & Signature:

Naznin Sultana Choity

20301275

Samiha Takmim

21301222

Abida Rahman

21201645

Abdullah Al Shaid

20201103

Md.Tanzim Hossain

21201131

Approval

The thesis/project titled "Benchmarking Vision-Language Models for traffic scene understanding in South Asian Traffic Environments" submitted by

1. Naznin Sultana Choity (20301275)
2. Samiha Takmim (21301222)
3. Abida Rahman (21201645)
4. Abdullah Al Shaid (20201103)
5. Md.Tanzim Hossain (21201131)

of [Spring], [2025] has been accepted as satisfactory in partial fulfillment of the requirement for the degree of B.Sc. in Computer Science in Summer 2025.

Examining Committee:

Supervisor:
(Dr.Swakkhar Shatabda)

Dr.Swakkhar Shatabda

Professor
Department of Computer Science and Engineering
BRAC University

Co-Supervisor:
(Member)

Name of Co-Supervisor

Designation
Department
Institution

Thesis Coordinator:
(Member)

Dr. Md. Golam Rabiul Alam

Professor
Department of Computer Science and Engineering
Brac University

Head of Department:
(Chair)

Dr. Sadia Hamid Kazi

Chairperson and Associate Professor
Department of Computer Science and Engineering
Brac University

0.1 Abstract

A substantial body of research has investigated methods to promote safer driving, yet many challenges persist, particularly in environments with complex traffic patterns. This study focuses on how drivers evaluate their surroundings and make safety-critical decisions, with a specific emphasis on the role of deep learning in enhancing safe driving. Deep learning enables rapid identification of hazardous situations by providing real-time feedback and alert mechanisms, thereby improving driving behavior and reducing risks.
The objective of this work is to develop an AI-powered driving assistance system designed to enhance road safety, especially for inexperienced drivers. Real-world driving conditions were incorporated by collecting YouTube footage across diverse road types and traffic densities. Experts annotated the dataset by labeling key objects and identifying context-specific risk factors, decision-making cues, and hazard indicators.
The system is powered by a custom-designed AI model capable of providing context-aware guidance, regulatory reminders, and hazard alerts in real time. By leveraging expert-annotated data and multimodal deep learning techniques, the system delivers personalized and immediate support to increase driver situational awareness, confidence, and safety.

keywords: Safe Driving, Vision-Language Models, Deep Learning, Driving Behavior, Road Scene Understanding

Acknowledgement

I express my deepest gratitude to my supervisor, Dr. Swakkher Shatabda, Professor, Department of Computer Science and Engineering, BRAC University, for his invaluable guidance, continuous support, and insightful supervision throughout this research. His expertise, constructive feedback, and encouragement have been instrumental in shaping the quality and direction of this work. I also extend my sincere appreciation to the Department of Computer Science and Engineering, BRAC University, for providing the academic environment, resources, and facilities necessary for completing this study.My heartfelt thanks go to my friends and peers for their helpful discussions, suggestions, and motivation. Finally, I am profoundly grateful to my family for their unwavering love, patience, and support. Their encouragement has been my greatest source of strength.

— Naznin Sultana Choity
BRAC University
November 2025

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

This chapter introduces the research by discussing the importance of safe driving, the relevance of Vision-Language Models (VLMs) for understanding traffic scenes, and presents the motivation, objectives, methodology, scope, and limitations of the study. It ends by outlining the structure of the research, providing readers with a clear understanding of its purpose and justification. The introduction offers an overview that addresses the background, the problem statement, objectives, methodology, and challenges associated with the investigation.

1.1 Background

The origins of this research can be found prior to the growth of interest in automobiles and the obstacles that drivers without experience face on the road. Driving skills that are fundamental, such as changing lanes,taking over,sustaining a safe momentum, can occasionally be difficult for new drivers to grasp.These challenges resulting in driver collisions causes an overall decrease in traffic flow and an upsurge in driving anxiety. The adoption of autonomous vehicles, or AVs, in hazardous circumstances, such as total vehicle control, may assist in addressing driving issues. Automobile manufacturers have introduced functions such as Tesla's Automobility, but they are not yet able to reach a level where it will be completely reliable. The Tesla Autopilot technological advances demands where the driver's hands will be kept on the wheel at all times but it will also be initiating automatic braking when the driver leaves their hands away from the wheel. The present condition of contemporary roadways has been defined by an assortment of functional limitations, mainly appearing from the need for selecting the right steps under complicated operating situations, responding to uncertainty controlled by human and keeping up urban safety measures.Researchers are looking at ways to improve the practical use for autonomous vehicle technology.

Considering the large amount of accidents and fatalities documented yearly, road safety continues to be an important concern in Bangladesh. The Road Security Foundation estimates that during 2019 and June 2024, there had been 32,733 road occurrences, ultimately resulting in 35,384 deaths with 53,196 hospitalizations [17]. The absence of extensive information, and these specifically links novice drivers to crashes, emphasizes the necessity for additional studies in this domain, given that there are quite few particulars with regard to the amount of instances brought on by foolish or unfamiliar drivers. Due to the absence of data and a range of reporting requirements, calculating the number of crashes of self-driving vehicles (AVs) is a difficult undertaking. According to California statistics,

AVs participated in 150 collapses on 5.7 million km in 2022, which is equivalent to 26.3 incidents for each million km driven [13]. The United States' human-operated vehicle death toll, in the opposite conjunction, was 1.33 fatalities per 100 million miles of vehicle in 2022.
Based on an investigation that examines Waymo's autonomous ride-hailing service, the power source injury-reported accident rate for its cars was 0.41 incidents per million miles, which is nearly 80 percent less than the usual rate for human-driven motorbikes, which is 2.78 mishaps per million miles [17] [13] [12].

1.2 Rational of the Study or Motivation

The genesis of this research can be traced back to the increasing affection for automobiles and the challenges faced by new drivers while operating them. It is challenging for humans to acquire fundamental driving skills, including passing maneuvers, lane modifications, and sustaining secure speeds. Driving tension is exacerbated by these issues, which also result in road accidents and a reduction in traffic efficiency. By assuming complete control of vehicles, AV technology enables the resolution of operational difficulties in critical scenarios. Tesla's Autopilot technology has been integrated into the automotive industry; however, these self-driving systems must achieve complete operational reliability. Drivers are required to maintain their hands on the wheel at all times while operating Tesla Autopilot systems, which is a critical requirement. The system's design dictates that automatic braking is initiated when the motorist ceases to grasp the wheel. The current traffic systems necessitate a variety of operational constraints to facilitate decision-making in complex driving situations, as well as to manage unpredictable human-controlled vehicles and maintain urban safety systems. The current focus of research is the development of novel solutions to enhance the efficacy of autonomous vehicle systems. Pedestrians as well as challenging and crowded terrain create major issues when it comes to navigation within diverse areas. Drivers operating in South Asian countries need to familiarize themselves with frequent traffic violations and poorly conceived road designs together with sudden pedestrian movements that characterize driving conditions there. Proper datasets become essential because they create accurate representations of complex real-life training conditions which autonomous systems encounter. The precision and reliability of vehicles in multiple driving environments will improve because data collection will now incorporate information about unpredictable road conditions with erratic traffic. Improving AV technology for real-world conditions requires enhanced data collection for both datasets and autonomous system programs.

1.3 Problem Statement

High-quality annotated driving data from BDD100K and Drama helps support the present training and testing procedures for autonomous vehicle systems. These datasets have made major progress in AV technology by documenting various forms of urban scenes in conjunction with different traffic scenarios and driving conditions. Besides structured driving environments most datasets handle today show limited capability to represent traffic conditions found within South Asian countries. Autonomous driving models face substantial challenges due to the unstructured road networks together with unpredictable pedestrian motions and frequent traffic violations that mark the South Asian driving condi-

tions. The three South Asian datasets (DhakaAI [32], BNVD [24], PoribohonBD [4]) require additional advancement to match the extensive capabilities of BDD100K and Drama assessments. The detection performance of the datasets receives improvements because of improved object tracking methods and segmentation algorithms and multi-sensor combination techniques that advance its operational capability. The incorporation of enhanced annotations into autonomous vehicle training leads to improved operational efficiency because detailed elements like traffic lane signals and roadflow distribution data alongside driving pattern information are included in custom annotations. These additions will turn the dataset into a robust data collection platform that simultaneously displays South Asian traffic behaviors and satisfies international research requirements within automotive vehicle fields. The methodology will enable worldwide market expansion for autonomously adaptable vehicles because of its reliability advancement capabilities.

1.4 Objective

Our first goal is to create a complete dataset of 1600 real-world traffic images to cover a variety of traffic conditions in Bangladesh and other neighboring countries, such as congested intersections, rural roads, and urban highways. In order to create this dataset, YouTube videos that capture road traffic with scenes in Bangladesh were gathered. Ten second lapses were taken from each video, and representative frames were chosen to record different traffic dynamics. This was followed by curation of these frames manually in order to be diverse and relevant. Lastly, 32 well structured questions were answered on each of the chosen pictures, which were developed according to the local traffic and road conditions in Bangladesh,another primary purpose of us is Model Benchmarking and Evaluation. To compare and contrast the performance of four state-of-the-art Vision-Language Models Qwen, Llama, Gemma, and LlaVa on the understanding and interpretation of complex and unstructured traffic scenes, as seen in South Asian settings. The analysis will be aimed at evaluating how well the models can identify, explain and reason chaotic or unconventional traffic behavior, therefore restricting their limitations and their possible application in intelligent transportation systems.

1.5 Methodology in Brief

The investigation will employ a variety of methods to enhance the efficacy of autonomous operation in response to a variety of external circumstances. The image detector monitoring activities in Bangladesh will result in the collection of a significant number of 17,326 car images. Subsequently, the Bangladesh Native Vehicle Detection (BNVD) database contains a total of 81,542 motor vehicle recognition entries. Qwen, LLaVA, Llama, Gemma these models were used to analyze the data set to identify automobiles, with versions v5 [5] through v8. Computers are utilized in conventional image processing techniques to identify vehicles and target targets. This technology enhances the motorist's comprehension of the system's precision efficiency and rapidity when operating in accordance with demanding options.In order to evaluate VLM Auto [20], an automated pipeline must be established between Robot Operating System 2 and the CARLA simulator [2]. As a component of their learning process, vision language algorithms generate driving instructions utilizing incoming visual signals. This makes real world investigations feasible. The integration of autonomous transportation and driver awareness has resulted in an

increase in both efficiency and reliability. We will assess the enhancement of the data set in comparison to existing datasets.

1.6 Scopes and Challenges

The research evaluates autonomous driving technology by combining technological assessment with ecological aspects alongside computational aspects which incorporates regional integration. Our ecological assessment analyzes both weather patterns and fluorescent lighting changes to establish Bangladeshi environmental conditions for vehicles. But nonetheless, the project foundation creates substantial difficulties for teams working on both data tagging operations and database information matching processes. The recognition system for ego lanes completes inspections simultaneously on highway illuminations with environmental surveillance of traffic environments. The system has performance problems when detecting lanes and identifying errors in areas with excessive noise. Several aspects of the environment generate challenges for lane identification system performance levels. Visual language models (VLMs) need many technical barriers to advance autonomous vehicles because of their development requirements. The two main hurdles to achieving practical system testing involve real world driving condition assessments and standardized testing requirements across various domains. Automated driving functions require VLM based predictions and decision-making procedures that surpass acceptable levels of computing power in autonomous vehicles. The quality standard of autonomous driving platforms should improve to boost speed and security until current problems receive solutions.

1.7 Key Learnings and Insights

Through the review of existing studies and related technologies, it has become evident that deep learning offers significant potential for improving driver safety and efficiency compared to traditional rule based systems. Neural network architectures such as CNNs and RNNs can effectively process visual and sequential data to recognize driving patterns, predict risks, and support decision making in real time. The study also highlights that driver safety and efficiency are closely interconnected, and optimizing one often contributes to the other. These insights have guided the direction of this research toward developing a robust and adaptive deep learning based framework for safer and more efficient driving systems.

Chapter 2

Literature Review

Large language models (LLMs), deep learning, and multifaceted information processing techniques have recently been employed in modern automotive systems to enhance safety, situational awareness, and decision making [1] [2]. NLP and vision-language modeling approaches have led to systems such as DRAMA [3], DriveGPT-4 [4], and LMDrive [5], which assess hazards, interpret complex road environments, and support real-time driving interaction. Technologies including LLaDA [6] and OmniDrive [23] further improve contextual understanding and responsiveness in dynamic traffic scenarios. Simulation frameworks like CARLA [8] are frequently combined with advanced vision language models such as VLM Auto [10] to achieve strong performance in scenario analysis.
Datasets such as NuScenes QA [23] strengthen 3D perception evaluation, while perception and control models like YOLOv8 [12], Soft Actor Critic (SAC) [13], and Model Predictive Control (MPC) [14] contribute to improvements in object detection, route planning, and obstacle avoidance. Complex driving behaviors can be explored through simulators such as SCANeR Studio [16] and CARLA [8], while generative and predictive algorithms including MiniSom [17] and LightGBM [18] assist in identifying anomalies and refining training data.
Despite these advancements, challenges persist in processing speed, data variability, and human machine interaction [19]. Improved language-based interfaces, greater reliance on edge computing [20], and stronger localized information processing capabilities are expected to guide future research. Ultimately, autonomous driving systems will continue to become safer, more accurate, and more practical as these issues are systematically addressed.

2.1 Preliminaries

The current state of autonomous driving has made progress through sophisticated safety mechanisms developed for operational efficiency and system adaptability. Basic risk assessment methods, together with multi mode data processing and natural language processing of core elements, achieve better situational awareness and driving decisions during dynamic road conditions. The DRAMA driving risk assessment mechanism alongside Large Language Models exemplifies a system base build from language that enhances risk enhances risk assessments throughout human machines communication.
The meaning of the application of self driving cars with advanced capabilities Autonomous Driving, or AD does not imply driving safely but instead provides the reason why the cars are moving in the first place. The existing self driving programs seem a black box in

which one cannot peep in or examine their decisions. AD must be readable in such a way that it becomes a safe and reliable one.
The Multimodal Large Language Models (MLLMs) is the solution to this query. They include AI models such as those that we test, able to process an image or a video, and learn and text generate. This may be said to imply that it is not only receiving what is going on the road that an MLLM is able to provide a user with an awareness of what is causing an object to turn or stop.
We establish how accurate various significant MLLMs are in justifying their decisions in our study. We take the best business model Qwen and compare it to some existing open-source models LlaVa [9], Llama [15] and gemma [26] to research what type of AI is the most efficient on such an extremely important task as safety.

2.2 Review of Existing Research

In [24], Saha et al. proposed introducing the Bangladesh Native Vehicle Dataset (BNVD), that contains 81,542 assigned situations, 17,326 photographs, including 17 automobile categories, in the book "Bangladeshi Native Vehicle Detection in the Wild" The database overcomes the absence of region-specific data needed for recognizing vehicles by documenting an assortment of circumstances notably changeable illumination, the climate, and vehicle rotations. YOLO v5 to v8 were employed to test BNVD; YOLO v8 achieved the best (0.848 mAP at 50 percent IoU).The collection addresses the dearth of local data necessary for vehicle identification by documenting a variety of instances, including variable electricity, the environment, and automobile orientation. YOLO v5 to v8 have been employed to test BNVD; YOLO v8 performed the best (0.848 mAP at 50 percent IoU). The dataset revealed 0.931 mAP in sunlight and 0.867 mAP in the evening, exhibiting excellent performance under various situations. Class disorder, low precision for tiny automobiles, annotation problems, and computing limits endure as difficulties, nonetheless. This research recommends increasing the variety of information sets, mixing classes, expanding annotation employing automated techniques, expanding low light detection, and optimizing instantaneous fashion release ways for beneficial ITS uses in Albania with the objective for improved performance.
In [29], Xu et al. analyzed the shortcomings of current systems for autonomous driving with regard to rapidity, precision in hazard recognition, and smooth routing in rapidly changing and complex road conditions. The study examines the challenges connected with autonomous vehicle operation, such as having to perform precise and quick obstacle recognition, successful planning of routes, and distance measurement. The YOLOv7x CM model's improved detection feature makes use of advanced techniques like CBAM and MPDioU to enhance target identification efficiency and precision. A binocular distance calculation algorithm is meant for removing oversights brought on by off center angles. The TEB-S algorithm features boundaries for accomplishing effective and effortless obstacle avoidance
In [1], Berrie et al. represented a real time vision based system which completes ego lane analysis through departure scenario detection and adjacent lane detection and marking type identification throughout lane change and estimation periods. Monocular images receive Hough lines along with Kalman filtering to perform lane detection while particle filtering with spline models calculates the curvature. The system achieves better feature extraction with Inverse Perspective Mapping (IPM) while performing classification of lane markings and simultaneously detects road signs and crosswalks and detects adjacent

lanes objects. The newly created set of more than 15,000 video frames showcased diverse road situation scenarios. The detection system accompanied by super fast processing speed surpassing 30 FPS makes ELAS suitable for real time applications in autonomous driving systems and ADAS technology.

In [8], Lee et al. suggested lane detection for advanced driver assistance systems and autonomous driving.This study focuses on enhanced driver assistance and autonomous driving technologies. Lane detection is required for all autonomous car functions.This research relies on destination notifications and path positioning assistance technologies. The authors found many constraints in standard computer approaches. According to studies, computer vision methods have several execution problems. It tries to improve detection accuracy while maintaining system stability.The project aims to improve deep learning systems' road lane recognition and travel prediction in different climates.The authors built DSUNet, a lightweight architectural variant of UNet, for lane detection.Lane detection and route prediction are high quality capabilities of this technology. This variant relies on depthwise separable convolution.Deepooled separable convolution speeds model deployment and reduces model size. They built their system utilizing DSUNet and route prediction algorithms.The researchers created a simulation tool to test their system's real time operational efficacy using CNN-PP and their prediction algorithm.CNN's performance in various driving conditions. Steady state study showed that DSUNet runs faster with fewer models than other methods.This latest version of UNet improves speed and memory usage while maintaining dynamic traffic analysis precision.

In [3], Mohseni et al. suggested applying a Dynamic Obstacle Avoidance Model Predictive Control (MPC) technique for autonomous cars in unexplored regions. The primary objective is to enhance comfort and productivity while simultaneously avoiding static and dynamic challenges to safe operation. This was accomplished through the use of MPC and Deep Learning Interface. Deep learning projections collision probabilities utilizing a collection of artificial neural networks, while the MPC improves pathways based on restrictions.To encourage secure and effective travel, a velocity dependent accident charge has been determined by considering the possibility of a collision. The LiDAR sensor data is crucial for constructing collision cones and forecasting risk since it provides data regarding obstacle positions and velocity. This method has been evaluated in a laboratory setting to show the ability for adaptation to new traffic conditions and environment.

In [19], Elallid et al. proposed the objective of the project is to strengthen the capability of self driving vehicles (AVs) to move around in unfavorable traffic conditions, particularly when approaching and navigating roundabouts. It includes the difficulties of minimizing accidents along with making decisions in rush hour traffic. The system has its foundation on the Soft Actor Critic (SAC) algorithm and Convolutional Neural Networks (CNNs). The CNN extracts characteristics based on photos captured by the AV's main camera and merges them with direction vectors. These inputs allow the SAC model to render instantaneous fashion navigational predictions. The CARLA simulator is used to assess a company's approach and ensure realistic traffic environments.

In [23], Qian et al. introduced NuScenes-QA, the first large scale autonomous driving Visual Question Answering (VQA) benchmark, is discussed in the paper. It addresses the challenges of answering natural language queries about dynamic street view sceneries utilizing multimodal data like LiDAR point clouds and photos. The largest 3D VQA dataset has 34,000 visual sequences and 460,000 question-and-answer combinations. Current VQA datasets are lacking, which this work addresses. Due to their static, single frame, or inside focus, VQA datasets have been unsuitable for autonomous driving in dynamic ex-

terior contexts. NuScenes 3D perception dataset provides annotated 3D object bounding boxes. Graph nodes represent items (e.g., automobiles, people) and edges describe spatial connections (”front left”). These components depict visual situations. Using scene graphs to manually create question templates (e.g., ”Are there any moving cars to the left of the bus?”) produces different question-answer combinations. Five types of inquiries are existence, counting, query object, query status, and comparison. A spatial reasoning question VS a non-reasoning inquiry. Use BEVDet to extract features from photos. The spatial data from CenterPoint can be understood using LiDAR. LiDAR and image data are fused in MSMDFusion to allow multi modal reasoning. Conventional QA designs use MCAN (Transformer-based) or BUTD to decode the answer.

In [31], Zheng et al. proposed a graph based Visual Question Responding (GVQA) architecture and Vision Language Modeling (VLMs) are used to create a context-aware making choices platform for self driving cars in SimpleLLM4AD. This strategy rethinks the perception, prediction, planning, and behavior pipeline to reflect visual question-answering (VQA) efforts using an intuitive dependency structure. The continual flow of data between phases allows demanding thinking and boosts dependability in quickly changing and tough driving situations. This technique divides autonomous driving problems into four VQA stages: thinking (object identification), calculation (motion prediction), design (driving strategy generation), and behavior. It uses Graph VQA, where logical dependent relationships provide context for subsequent actions.The vision encoder is the image text trained InternViT-6B Vision Transformer. This language decoder allows context aware reasoning. The Graph of Thought (GoT) structure reduces inputs and improves contextual cognition by tokenizing nuScenes graphical components and coordinating them with written information to produce answers. Object detection improves decision making by include location, color, and orientation. The system scored similarly to earlier techniques on the DriveLM nuScenes dataset, with strong context-dependent cognitive benefits.

In [27], Wang et al. alleged big language models govern massive neural networks in OmniDrive, an autonomous car control software. OmniDrive Agent evaluates multi view pictures for the operating system using its vision language 3D paradigm. This lets individuals analyze multi view images while the system retains 3D environmental awareness. In the Omnidrive nuscenes benchmark, the system simulates planning and reasoning situations across three categories. The system uses the Q-Former 3D architecture for two dimensional pre training and three dimensional fine tuning. High resolution multi view input processing challenges can be solved with multi task learning and temporal techniques.Progressive planning uses Q Former 3D to blend LLMs with 3D visual data solution pathways. Dynamic learning modeling for visual input analysis integrated object recognition and management training on a single platform. Automation in evaluation. The benchmark evaluates counterfactual thinking during reasoned decision making operations planning. High resolution multi view imagery is processed better by two dimensional pre trained knowledge systems using three dimensional methods.

In [10], Malla et al. proposed a vehicle danger assessment mechanism with a caption module (DRAMA) that detects vehicle risks and communicates the danger to drivers through natural translation of words. The author generated a language based description dataset by data labeling films and objects to extract unique information and build a new study path. The annotation framework classifies driving scenarios, marks boxes bound objects at risk, and provides driver proposal inputs by synchronizing video, IMU, and vehicle control data. Data loader scripts split training and validation data and provide testing sets with word to index and index to word conversion mappings. The framework structure

improves situational awareness and decision making, improving danger location identification and natural language captioning for driving support technologies. In [28], Xu et al. DriveGPT4 either updated existing data or created new proprietary datasets instead of developing new ones. To increase the model's query processing capabilities, the author combined BDD X database with ChatGPT. The BDD X dataset comprised 20k video samples with speed, turning angle, and vehicle motion descriptions. The author created 56k samples by merging BDD X and ChatGPT inquiries because the prior dataset's questions were limited and standardized.

In [25], Shao et al. proposed LMDrive, a game changing closed loop technology that runs constantly from start to finish. Large language models (LLM) allow LMDrive to analyze complex situations and provide recommendations. LMLs guide operations and think and talk through challenging problems. By integrating complicated experimental data from sensors and verbal conversations, the system provides real time control signals. LMDrive processes spoken input better than traditional voice communication systems. The benchmark and data set create an evaluation framework for automatic driving based on language interfaces in real world operational settings. The original LMDrive approach uses natural language protocols to communicate between cars. The natural language base improves flexibility and interpretability. Independent autonomous vehicle systems have little operational flexibility. Every imaginable combination of conditions and unanticipated occurrences and their interhuman interactions between problem solvers is examined in the research. Advanced human machine interaction, clearer explanations, and stronger reasoning functions generate three key goals. Integrated systems for conversational reasoning.

In [21], Li et al. proposes LLaDA (Large Language Driving Assistant), a framework for adapting driving patterns and planning motion in new places with diverse traffic rules and traditions. LLaDA helps human drivers and autonomous automobiles (AVs) comprehend local traffic laws and adjust driving plans. It addresses two major issues: geo fenced AVs' inability to adapt to regional traffic law differences (e.g., "right turn on red" regulations in New York and San Francisco) and their inability to handle unexpected situations (e.g., honking or unusual obstacles) without training.LLaDA leverages already trained LLMs to adjust traffic restrictions in real time, enhancing mobility planning across regions (e.g., Singapore to Boston) and language-based logic for unexpected driving conditions. Natural language inputs like scene descriptions, local traffic restrictions, and unexpected events like animals crossing change driving policies. An LLM (e.g., GPT-4) updates rules, setting descriptions, and baseline driving plans to comply with local laws and discuss unexpected scenarios, while a Traffic Rule Extractor (TRE) extracts relevant legislation from local traffic handbooks. Through connection with AV systems like GPT Driver and real time traffic suggestions ("Do not turn right on red in NYC"), LLaDA encourages AV motion planning.

In [11], Mao et al. proposed an automated driving system called Agent Driver employs, Large Language Models (LLMs) to make decisions and reason like humans, according to the study "A Language Agent for Autonomous Driving". A Tool Library for processing environmental input, a Behavioral Recall for experiencing knowledge, and a Reasoning Engine for advanced planning and self analysis distinguish Agent Driver from typical perception prediction planning frameworks, which lack flexibility and common sense. This method improves on rule based systems and demo based learning by allowing more nuanced and adaptable responses to dynamic driving conditions. The perception prediction planning framework of traditional self driving systems sometimes lacks human drivers'

common sense and experiential knowledge. This limitation can impair decision making in complicated or unexpected situations. Using large language models (LLMs) in driving operations, the Agent Driver system overcomes autonomous driving problems. The system uses a Tool Library for data collection and processing, a Cognitive Memory for experiential knowledge, and a Reasoning Engine for chain of thought reasoning for task and motion planning and intuitive driving self reflection. The Agent Driver system redefines autonomous driving mechanics to adapt to changing traffic conditions. These methods need extensive parameter tweaks after conversion, making bulk deployment impracticable for common use cases.

In [20], Guo et al. characterizes VLM Auto as an autonomous driving assistant powered by Visual Language Models (VLM) which enhances vehicle responsiveness in complex driving scenarios. Traditional methods improve with VLMs which enable the interpretation of road situations for making decision choices. A Qwen VL model operates between VLM Auto and ROS2 while CARLA simulator provides the necessary connection to evaluate road conditions including weather elements and lighting together with present obstacles. The model training process operates on 221228 image samples together with their corresponding prompts which were sourced recently. The precision outcomes of VLM Auto reach 97.82 percent accuracy during CARLA simulations testing as well as real world conditions deliver 96.97 percent correct detections. Behavior adjustments based on visual cues within the system result in enhanced driving performance during simulated and actual operations through the combination of smooth movements with adaptable systems and generalized actions.

In [14], Teng et al. alleged the study "Motion Planning for Autonomous Driving: The State of the Art and Future Perspectives" aims to develop algorithms for autonomous cars that can adapt to complex driving circumstances safely and efficiently. We recommend pipeline and end to end planning. Pipeline planning modules are interrelated and focus on different planning aspects. Orchestrated tasks allow the system to make good decisions, examine the trajectory, and maximize the outcomes. Integrating training and validation to address all driving duties makes end to end planning a single learning problem. In [7], Hu et al. the UniAD architecture in "Planning oriented Autonomous Driving" presents end to end autonomous driving as a comparable method. Instead of separating observation and prediction from planning, the UniAD architecture integrates these roles to focus on planning. UniAD's query based nature relies on flexible searching methods to get contextual items instead of bounding boxes. The architecture enables elastic interaction models between several agents while reducing predicted errors from prior phases. In these publications, holistic, integrated techniques for autonomous driving systems emphasize task coordination and system efficiency. The latter approach promotes an all encompassing framework that emphasizes planning as a key driver of system function, whereas the former leverages separate organizational arrangements.

In [18], Chen et al. illustrated the revolutionary potential of end to end automated modes of transportation, while additionally recognizing the constraints of protection, generalization, and clarity. A combination of approaches that combines the positive aspects of both end to end and modular systems, as well as improvements in data utilization and simplicity, has the potential to expedite the development of safer and more trust worthy technology for autonomous vehicles.

In [19], Elallid et al. intends to enhance self driving vehicle (AV) mobility through challenging traffic scenarios including those leading to and within roundabouts. The project involves two challenges which are reducing accidents and making choices during peak

hours. The system builds upon SAC algorithm with CNNs as its core components. A CNN system utilizes the main camera photo images from the AV to extract characteristics which it combines with direction vectors. SAC models make immediate fashion related route predictions because they processing these inputs. The carla simulator serves assessment purposes to evaluate corporate driving approaches while generating authentic traffic simulation environments.

In [16], Wang et al. Demonstrated using real world transportation scenarios:There are two model phases here .Firstly,it learns traffic laws and then it constraints in order to improve efficiency to understand inert backdrops moving objects.Secondly it learns so that it can predict difficult situations which will make a similar video prediction to produce driving scenarios.Diffusion models are used so that it can provide authentic recordings and pictures to build the model of real driving films and driving data.Variational autoencoders (VAEs) are also used in that so that it can compress and recreate data for video production to improve the quality. Bev control, Magic Drive, and DrivingDiffusion have been used to guide picture generation.The movie will be made step by step using an auto regressive model so that it can forecast each component. Flow based models were used to simplify data. Lastly, two AIs can be used to compete using the generative adversarial networks so that one can make it and the other can compromise it.Structured traffic regulations ,narrative descriptions and circumstances are used by Drive Dreamer in order to anticipate the video instead of past video frames.Future road constructions will be predicted in Action Former in a format being compressed.These are the requirements of Auto Dm which will create video.The time conditions like day and weather will be changed in travelling conditions by text prompts.

In [22], Liu et al. implied studies methods for enhancing drivers security and ease when using intricate underground interchanges. Utilizing SCANeR STUDIO's 3D modeling and driving simulation, the study examined the effects of wall colors, speed bump creates, and signs on motorists behavior. The LightGBM machine learning model was used to analyze data from 24 participants, and it correctly categorized safety and comfort levels with a 97.06 percentage accuracy. The main results highlighted the importance of strengthening approach signage, braking zones, and visual cues for better driver guidance and decreased tension. The study presents an extensive foundation for integrating Simulation and machine learning methods into transportation planning, offering useful knowledge for the establishment of environmentally conscious and safer urban mobility systems.

In [30], Yun et al. employs a machine learning-enhanced approach to detect abnormal driving behavior in autonomous vehicles using simulations of three key scenarios: This study uses machine learning to detect abnormal driving situations occurring from sensor failures and vehicle collisions and counter flow traffic movements. Training data originated from Basic Safety Messages (BSMs) which were simulated through Vehicle to Everything (V2X) communication using the CANoePro tool. The researchers applied One Class SVM and K Means clustering followed by HDBSCAN and MiniSom to process vehicle data consisting of GPS coordinates with speed parameters and heading values. The MiniSom algorithm demonstrated exceptional performance by delivering complete accuracy throughout all checked scenarios. A proposed enhancement in this study added the linePosition parameter to BSM structure data to improve anomaly detection accuracy. Real time computations took place through an edge computing system that replicated a Roadside Unit to provide instant detection capabilities against abnormal vehicle maneuvers. The performance assessment of MiniSom demonstrated reliability by using accuracy

along with recall and precision measurements to validate its detection capabilities.
In [6], Malla et al. presumed the research uses basic computer vision techniques to produce a cost effective and efficient autonomous car system. Lane boundaries are established using a multi stage process that begins with color space transformation. Next is Canny Edge Detection, then Hough Line Transformation. Following this, vehicle steering is implemented. The project substitutes LiDAR sensors with inexpensive Raspberry Pi and Nvidia Jetson Nanos for operational efficacy and realistic outcomes. This vehicle's object detecting software activates automated brakes when it senses things ahead. The study team built the autonomous framework using a commercial remote controlled car. Vehicle automation technologies and cost effective monitoring methods have been developed to enable scientists to design accessible next generation self driving systems.
In [3], Mohseni et al. proved that autonomous cars should employ Dynamic Obstacle Avoidance Model Predictive Control (MPC) for driving in unknown terrain. The system aims to boost comfort during operations with productivity gains by managing obstacles that present safety risks during both static and dynamic tasks. The achievement became possible through MPC combined with Deep Learning Interface. A collection of artificial neural networks in deep learning generates collision probability projections while MPC implements path improvements through operational restrictions. An accident fee that depends on velocity exists to promote risky yet secure travel. The LiDAR sensor data serves as the foundation for constructing collision cones while generating risk forecast because it reveals information about obstacle positions and obstacle velocities. Researchers applied this technique to laboratory testing for evaluating its capacity to adjust to new traffic situations and environmental elements.
In [29], Xu at al. identified gaps in self driving technology systems between their speed and detection accuracy and route handling abilities in changing difficult traffic conditions. The research looks into three specific obstacles autonomous vehicles face during operation including fast obstacle detection and precise planning of routes along with effective distance measurement. The YOLOv7x CM model implements advanced techniques CBAM and MPDioU for its detection feature to boost target identification speed and precision. The binocular distance computation system functions to eliminate alignment issues that occur due to notched positions. TEB-S algorithm presents operational constraints specifically designed to help users avoid obstacles effectively without effort.
In [22], Liu et al. investigated techniques to boost automotive comfort and safety during complex situations in underground facilities. Research performed by SCANeR STUDIO employed its 3D modeling and driving simulation to study the behavioral effects of wall colors along with speed bump creates and road signs. A compilation of data from 24 participants underwent LightGBM machine learning analysis which achieved 97.06 percent accuracy for classification of safety and comfort categories. Results demonstrated that approach signage strength and braking zones and visual signals matter for appropriate driver guidance while reducing driver tension. The research establishes a full framework to merge Simulation approaches and machine learning methods with transportation planning which gives strong knowledge to develop advanced sustainable urban mobility systems.
In [30], Yun et al. developed a machine learning solution that detects abnormal driving actions for autonomous vehicles through three important scenarios: This study incorporates machine learning algorithms to identify abnormal driving conditions triggered by sensor failures and vehicle accidents and movements across traffic directions. Basic Safety Messages (BSMs) served as the source data which the CANoePro tool simulated through

Vehicle to Everything (V2X) communication channels. Scientists conducted data processing on vehicle data containing GPS coordinates alongside speed parameters and heading values through the application of One Class SVM and K Means clustering together with HDBSCAN and MiniSom. Every tested scenario delivered complete accuracy when the MiniSom algorithm performed its operations. The current research proposed a structural change to BSM structure data by implementing linePosition as a new parameter to enhance detection accuracy. The edge computing system conducted real time calculations for abnormailty detection purposes against vehicle movements by mimicking Roadside Unit functionality. Performance assessments of MiniSom used accuracy alongside recall and precision measurements for validating its detection capabilities.
The paper, DriveGPT4: Interpretable End to end Autonomous Driving via Large Language Model, is the most helpful study that proves this fact. The most appropriate example of demonstrating that MLLMs can be a self driving system is this paper because it reveals that MLLMs have the means. DriveGPT4 proved it could:
1. Looking at low level controls that are needed to control the car (steering, speed).
2. Provide extensive, human, justification to such actions (explaining).
But, as we read the paper, and other papers of the same kind, we discovered that one fallacy is present; it is extremely difficult to prove the soundness of the explanations. The existing procedures, which are employed in research, are not noteworthy due to the following reasons: Unstable Testing: It is often faced by researchers to test the quality of an explanation given by another AI with a single AI such as ChatGPT. This should not be trusted because the scoring AI will not be without its bias and every time it will not have the same outcome. Reliance on the AI: Here, the datasets during the training of the AIs may occasionally be generated or changed by various AIs. This will mean that the model will be learning to hallucinate or pretend to come up with imaginary facts that though they may seem correct, they are actually not.

2.3 Summary of Key Findings

Natural language processing and DriveGPT4 computer vision software enhance driving risk assessment and road safety, providing improved protection for autonomous vehicles. Modern autonomous driving systems benefit from advanced weather analysis, enabling better driver decision making. However, current systems lack effective risk communication, highlighting the need for more robust approaches. The operation of advanced autonomous vehicles relies heavily on research based strategies, integrating verified data to ensure accurate risk alert systems. User involvement further improves risk detection and system performance. Future research should focus on advanced risk evaluation algorithms, data integration, and enhanced user interfaces. Effective risk management is critical, as it will shape the design of next generation autonomous systems.
Based on our research, we conclude that while MLLMs can “talk” and process driving scenarios, there is no reliable method to evaluate the correctness of their responses. This thesis addresses the research gap: there is no standardized approach to test models like Qwen, Llama, Llava and Gemma against true expert responses.
To solve this, we propose a two-part evaluation system:
Ground Truth: Human experts annotate images and encode correct values in a structured JSON, creating an immutable reference document.
Objective AI Evaluation: Open-source AI checkers Gemma, Ollama, Qwen compare model outputs against the JSON ground truth. This removes human bias and scoring

variability, providing a precise, quantifiable measure of each model's reasoning accuracy. This approach establishes a systematic, unbiased framework for evaluating MLLMs in autonomous driving contexts.

Chapter 3

Requirements, Impacts and Constraints

This chapter describes what was required to be able to conduct the project successfully. It talks about technical requirements, social and environmental, ethical factors, standards to be followed, project management strategy and project risks. Through this chapter, the reader can get acquainted with the practical constraints, difficulties, and liabilities of the design of a safe and ethical AI based driving system.

3.1 Final Specifications and Requirements

The second important functional requirement of this project is the ability to identify and read all relevant aspects in typical South Asian road scenes (vehicles, pedestrians, non-standard traffic behavior, and signage) with sophisticated Vision-Language Models such as Qwen, Llava, Llama, and Gemma. The goal of the benchmarking framework is to evaluate the extent to which these models are able to reason and explain complex, chaotic or unconventional traffic events. Other non functional requirements are that the use of public video material must ensure data privacy, the annotation process must be efficient and scalable, and bias mitigation must be ensured, such that model performance is fair and robust regardless of the types of scenes and real world conditions.
The limitations in this work are due to the lack of locally representative labeled data, variable quality of sources of the images because of using publicly available footage, and compliance with regulatory aspects of copyright and data privacy. The models must prove to be useful without having to undergo a lot of retraining on the South Asian data, and the implementation must not conflict with the modern regional infrastructure to support intelligent transportation applications.
The solution must be in accordance with the National AI Strategy of Bangladesh that requires ethical application of data, high privacy standards and disclosure of automated decision making tools, particularly in transportation. Other standards and codes of practice are applicable IEEE and ISO standards on intelligent transportation systems and AI, local ITS master plans, and accepted annotation standards such as YOLO or MS COCO to promote interoperability and scalability in the future. The involvement of the stakeholders such as local transportation experts and regulatory authorities is also necessary during the project to ensure that the project is in line with the real life needs, and standards, goals, and functional needs are continuously monitored to improve the project.

3.2 Societal Impact

The application of Vision Language Models (VLMs) to analyze traffic scenes in Bangladesh and other South Asian countries has significant societal, health, safety, legal, and cultural implications. Socially, these technologies can enable more efficient, reliable, and equitable traffic management by providing real time updates on congestion, optimizing traffic flow, and enabling rapid responses to accidents or hazardous situations. This can reduce commuting times, improve convenience, boost economic productivity, and enhance overall urban quality of life.

From a public health perspective, AI driven scene understanding can address the disproportionately high rates of road deaths and injuries, particularly among vulnerable pedestrians and cyclists. Dangerous behaviors such as driving wrong way, lane violations, and oversized vehicle throughput can be automatically detected, allowing timely warnings and interventions that reduce accidents and enhance safety. Smart traffic systems can also help law enforcement allocate resources more effectively, monitor enforcement performance, and hold traffic managers accountable. However, the effectiveness of these benefits depends heavily on the quality and local relevance of the data used to train these models, as well as their performance in complex, often chaotic, regional road environments.

Legally, such technologies must comply with national data privacy laws and emerging AI regulations. While Bangladesh has yet to develop specific AI legislation, government policy emphasizes ethical information use, transparency, and safe AI deployment in transportation. Key considerations include the lawful acquisition and storage of video data, copyright protection, and ensuring AI models do not perpetuate biases that discriminate against certain groups.

Culturally, VLMs must be adapted to local conditions, including diverse vehicle types, informal traffic behaviors, and region-specific practices prevalent in South Asian streets. Models should recognize common local traffic participants such as auto-rickshaws, cycle vans, and buses and reflect local traffic patterns rather than Western norms. Furthermore, successful deployment requires building stakeholder trust, fostering public acceptance, and implementing systems that respect local values and expectations.

In summary, the introduction of Vision Language Models represents a transformative step toward smarter, safer, and more responsive transportation systems in South Asia. Their success depends on careful adaptation to local contexts, adherence to ethical and legal standards, and prioritization of social welfare, public safety, and health.

3.3 Environmental Impact

The successful implementation of any project, particularly the one that relates to intricate AI models and real life applications like the vision language model benchmarking used to analyze the traffic, is impossible without effective verbal and written communication with stakeholders. Documentation in the form of regular notes and journals assists in the recording of insights, decision making, challenges and progress. These records are a good resource of continuity and transparency of the project.

Report and project deliverables must be made at different stages clearly stating the objectives, methodology, findings and recommendations. These reports should be formatted effectively, succinct and audience driven; between technical experts and local authorities. Simple sentences and graphic materials like charts, tables, and images are used to improve comprehension.

Presentation is an important tool of summarizing the progress and the findings. To make meaningful oral presentations, it is necessary to plan the content, highlight the most important points that can be of interest to the stakeholders, and rehearse effective and clear presentation. Practical value may be transferred through demonstrations of the benchmarking system or AI capabilities, which can help address gaps between technical development and the expectations of stakeholders.
Journals, reports, presentations and demos, combined, promote collaboration, keep in line with stakeholder needs and enable informed decision-making during the project lifecycle.

3.4 Ethical Issues

During the solution design process, the ethical standards and academic duties must be vigorously followed. Some of the ethical concerns are related to the privacy of the data used, consent when the public videos are used, biases in model creation that might disproportionately impact some population groups, and responsibly reporting the limitations of the model. Academic integrity is ensured by transparency of the method, recognition of previous work, and lack of plagiarism. The professional responsibility demands reporting the results honestly and informing the stakeholders about the progress and the challenges. The design must also take into consideration the effects on society, in the sense of giving back to the community by improving the accessibility and safety of the traffic. These are practices that develop trust and credibility of the project.
All these competencies together guarantee that the project is planned, cost effective, and ethical in terms of providing effective and responsible development of solutions.

3.5 Standards - if applicable

We maintained several important standards to make sure it is reliable, safe, and ethical. All data collection and processing follow general privacy and data protection principles, according to the ISO/IEC 27001 standard for information security. The system design also considers ISO 26262, which focuses on the functional safety of road vehicles, to make sure that the model promotes safe driving behavior and does not create any new risks. In addition, the software development and testing process follow IEEE 829 standards to ensure proper documentation, testing and validation of results. By maintaining these standards the work ensures that all processes are safe, consistent, and meet professional and technical requirements.

3.6 Project Management Plan

The Project Management Plan will project the step by step procedure of implementation of the project to its completion. It involves the definition of project goals, specification and establishment of deadlines of every activity of work including data collection, model training, evaluation and reporting. The plan outlines the distribution of resources, such as the assignment of personnel, hardware and software and budget. It also sets milestones and deliverables against which to measure progress. To achieve transparency and alignment of project with the expectations of stakeholders, communication strategies as well as quality assurance methods are included.

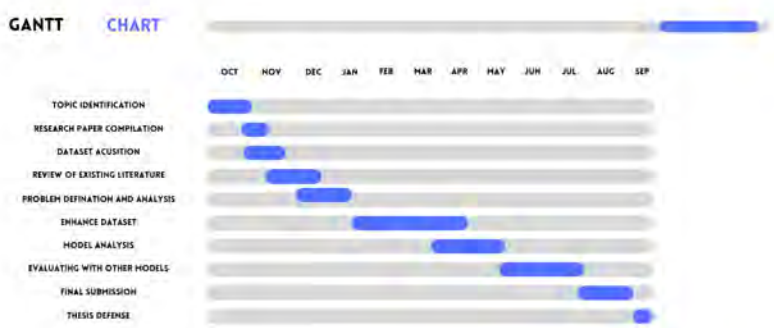

Figure 3.1: Gantt Chart

This chart 3.1 illustrates the planned period and budget for the driver safety system development. From initial planning to creation, evaluation, and final examination, it defines the sequence, length, and overlaps of the major project phases. This schedule helps in monitoring progress and organizing tasks for sure timely completion.

3.7 Risk Management

Risk Management is about uncovering any possible risks which can adversely affect the project like data quality, technology, time and budget overruns. The risks are measured in accordance with the probability of occurrence and their consequences and mitigation strategies are established to counter them in advance. Some of the strategies are data validation processes, fallback model selection, buffer time designations, and periodic progress review. The contingency plans will provide the ability to change the project in response to unexpected difficulty to keep the project on course without losing the resources used.

3.8 Economic Analysis

The project considers the economic feasibility and cost benefit aspects of implementing a safe driving system. The main costs include data collection, model development, software and hardware resources, and testing. These costs are weighed against the potential benefits, such as reducing road accidents, minimizing vehicle damage, and saving lives, which provide significant social and economic value. By evaluating both costs and benefits, the project shows that investing in this safe driving system is not only technically feasible but also economically justified, making it practical and sustainable for real world use.

Chapter 4

Proposed Methodology

The process of development of the dataset and benchmarking system explained here. It describes a stepwise description of image collection, generation of questions, validation, annotation of data, and assessment of Vision Language Models. The entire workflow, or raw video to final model predictions, is also presented in the chapter. It gives a clear picture on the conduct of the research in such way that it can be used by other people to replicate or advance the approach.

4.1 Corpus Creation

To support this research a dedicated corpus was created that focuses on real world traffic scenes and their understanding through visual question answering. The creation process followed a systematic approach to ensure the dataset was diverse, meaningful, and accurately represented real driving environments. The entire pipeline involved four major stages: data collection, question formation, validation, and annotation. In the first stage, we collect some raw videos from YOUTUBE and locally recorded dash cam footage. Then we extract screenshots within a 10 second interval. Several traffic scene screenshots were collected. Those videos were publicly available driving datasets such as BDD100K, Cityscapes, and Mapillary. The selection process aimed to capture a wide range of traffic conditions, including different weather types, lighting variations, and road settings such as highways, intersections, and city streets. This diversity helped the dataset reflect realistic and challenging driving situations. Once the images were collected, natural language questions were generated for each screenshot to evaluate a model's ability to interpret and reason about traffic scenes. The questions were designed to cover several aspects, including object recognition, action prediction and causal reasoning. These questions were taken from many research papers also created based on general driving knowledge and traffic behavior patterns to ensure their relevance and realism. To ensure the quality of the dataset, each question underwent a validation process. A traffic police examined all questions to verify that the questions were meaningful, clear and could be logically answered using the given visual information.
Any question that was ambiguous, grammatically incorrect or not grounded in the image context was either modified or discarded. This validation ensured the dataset's consistency and usefulness for traffic scene understanding. In the final stage, annotations were added to each image. Important objects such as vehicles, pedestrians, traffic lights and lanes were marked with bounding boxes. The answers to the questions were also annotated manually. This resulted in a well structured and high quality dataset consisting of

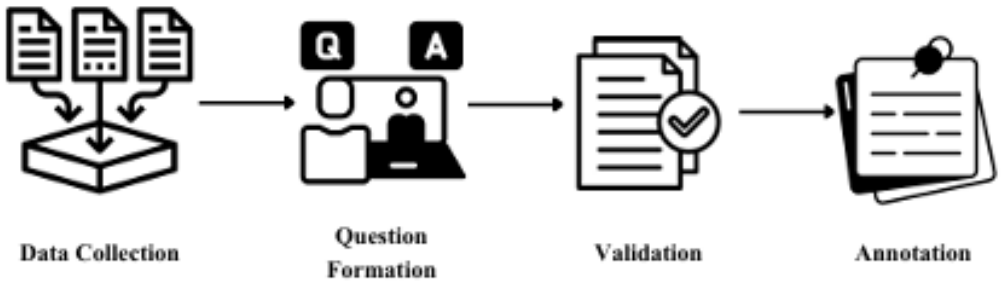

Figure 4.1: Corpus Creation

question answer pairs for each traffic scene. Furthermore, we have everyone of the five participants manually examine the entire dataset. During this entire process, we found that approximately forty percent of the inquiries had differences in viewpoint between the members. In a few cases, two members assigned an answer as erroneous while three members classified it as correct. For these types of disputes, we adopted the majority vote technique, indicating the final choice was based on what the majority agreed on. This contributed to ensuring that the confirmed information was as precise and fair as possible. Finally, the entire dataset creation process can be represented as a pipeline consisting of the following steps: data collection $\rightarrow$ question formation $\rightarrow$ validation $\rightarrow$ annotation.

4.2 Vision-Language Models Used in This Study

This study evaluates four modern Vision-Language Models (VLMs) capable of interpreting traffic scenes through combined visual and linguistic reasoning. These models were selected because they represent different design philosophies, parameter sizes, and computational demands, while being easily accessible through the Ollama framework. By testing all models on the same curated South Asian traffic dataset, we aimed to understand how well current VLMs handle unstructured, culturally diverse road environments. The models used in this research are:

- **Llama 3.2 Vision (11B)**
- **LLaVA-Phi3 (3.8B)**
- **Qwen2.5-VL (7B)**
- **Gemma 3 (4B)**

Llama 3.2 Vision (11B)

Llama 3.2 Vision is Meta's multimodal extension of the Llama 3.2 series, designed to process high-resolution images and generate coherent natural language responses. With 11 billion parameters, it is the largest model evaluated.

Strengths:

- Strong language reasoning and detailed explanations.
- Good extraction of visual information from structured scenes.
- Performs well on general visual question-answering tasks.

Limitations:

- Struggles with chaotic, unstructured South Asian traffic.
- Often misinterprets culturally specific driving patterns.

Reason for inclusion: Serves as a high-capacity benchmark for evaluating how a large foundational multimodal model handles complex traffic scenes outside its training distribution.

LLaVA-Phi3 (3.8B)

LLaVA-Phi3 is the smallest and fastest model in our benchmark. It integrates Phi-3's language capabilities with LLaVA's visual alignment, resulting in a compact 3.8B-parameter system suitable for lightweight applications.

Strengths:

- Very fast and computationally efficient.
- Performs well on simple and direct questions.
- Can run effectively on lower-end hardware.

Limitations:

- Limited reasoning depth due to its small size.
- Performance drops in dense or visually complex scenes.

Reason for inclusion: To assess whether lightweight VLMs can handle real-time reasoning scenarios typical of safe driving applications.

Qwen2.5-VL (7B)

Qwen2.5-VL is Alibaba's 7B-parameter model optimized for grounded visual reasoning. It demonstrated the strongest contextual understanding and overall accuracy among the tested models.

Strengths:

- Excellent visual grounding and contextual interpretation.
- Handles complex, cluttered scenes effectively.
- Robust across varying traffic environments.

Limitations:

- Slower than compact models due to larger architecture.
- Occasionally produces overconfident incorrect answers.

Reason for inclusion: Its balanced size and strong reasoning make it representative of practical mid-sized VLM deployments.

Gemma 3 (4B)

Gemma 3 is Google's lightweight multimodal model emphasizing stable and predictable behavior. With 4B parameters, it offers a middle ground between performance and efficiency.
Strengths:

- Consistent predictions across different scene types.
- Good balance between speed and reasoning capability.
- Rarely produces extreme hallucinations or incorrect outputs.

Limitations:

- Does not reach the higher accuracy levels of larger models.
- Struggles with highly congested or irregular scenes.

Reason for inclusion: Provides insight into how smaller but stable VLMs adapt to real-world traffic understanding tasks.

Model Comparison Summary

Model	Parameters	Key Strength	Main Limitation
Llama 3.2 Vision	11B	Strong reasoning ability	Weaker in chaotic scenes
LLaVA-Phi3	3.8B	Fast and efficient	Limited reasoning depth
Qwen2.5-VL	7B	Best contextual understanding	Overconfident errors
Gemma 3	4B	Stable and consistent	Lower peak accuracy

Table 4.1: Comparison of the Vision-Language Models used in this study.

These four models together provide a comprehensive perspective on current VLM capabilities, ranging from small and efficient systems to larger models capable of deeper reasoning. Their comparative performance forms the basis of the evaluation presented in the next section.

4.3 Design Process or Methodology Overview

The overall methodology of this research was designed to develop and evaluate a visual question answering model specialized for traffic scene understanding. The goal of the system is to take a traffic scene image and a related natural language question as input, and then generate an appropriate text based answer as output. This process allows the model to interpret visual elements, understand the context of the question, and reason about real world traffic situations.
The experimental framework consists of several major components: data preprocessing, feature extraction, question encoding, multimodal fusion, and answer generation. Each component plays an essential role in ensuring accurate and meaningful predictions.In the data preprocessing stage, the collected traffic scene images were resized and normalized to maintain consistency. The questions were cleaned by removing unnecessary symbols,

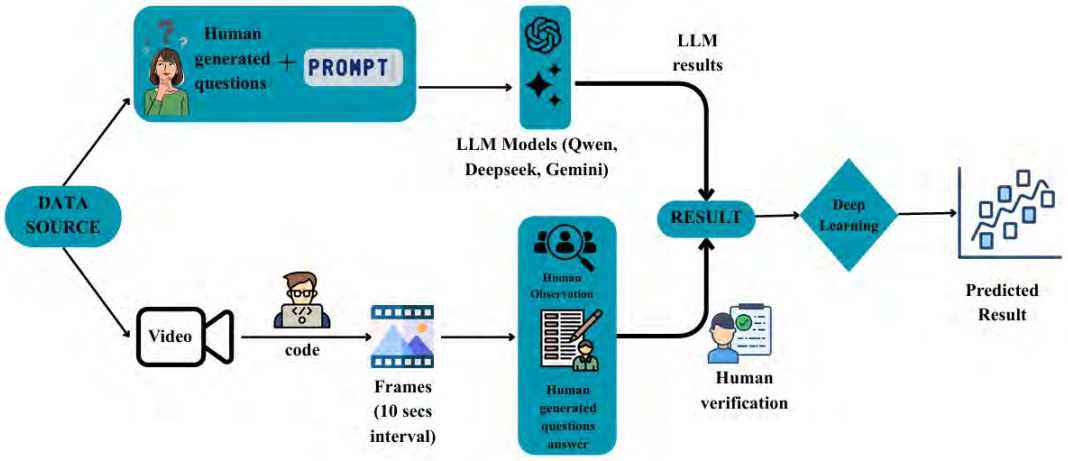

Figure 4.2: Pipeline for Benchmark Dataset Creation and Model Evaluation

converting text to lowercase, and tokenizing the words. Both images and text inputs were then transformed into vector representations suitable for deep learning models.

For feature extraction, a pretrained convolutional neural network (CNN) was used to extract high level visual features from each image. These features capture important details like vehicles, pedestrians, and road signals. For textual input, a recurrent neural network or transformer-based encoder was used to convert the question into a dense feature vector that captures its semantic meaning.In the multimodal fusion stage, the visual and textual features were combined to form a joint representation. This step enables the model to connect the question with relevant visual information from the image. For example, if the question asks, " are there any pedestrians crossing the road?"The fused representation helps the model focus on the human regions in the image.

After fusion, the answer generation module predicts the most suitable textual answer. This can be done through a classification layer (for short, predefined answers) or a sequence generation model (for descriptive answers). The model output is a text response that directly answers the given question based on the visual context.During the training phase, the model was optimized using cross-entropy loss between predicted and ground truth answers. Different hyper parameters such as learning rate, batch size and optimizer type were tuned experimentally to achieve the best performance.

Overall, the entire system can be summarized as a process where the model receives an image question pair as input and generates a textual answer as output.

This diagram 4.2 protraits a "DATA SOURCE" element labelled "Video" is connected to a component that gathers frames at 10 second periods using code, converting the initial video footage into analyzable pictures. In the block above, an individual generates "Human generated questions + PROMPT," these are plugged into an ensemble of LLM models, namely Qwen, Deepseek, and Gemini, to create "LLM results." In the center, a "RESULT" section demonstrates what people see and "Human generated questions answer," followed by "Human verification," indicating why people investigate and examine the model conclusions before they move next. On the right side, these demonstrated data points are fed into a "Deep Learning" module, finalizing in another "Predicted Result" segment which displays the model's output.

This graphic 4.3 shows a workflow model of an LLM processor pipeline. On the left, an

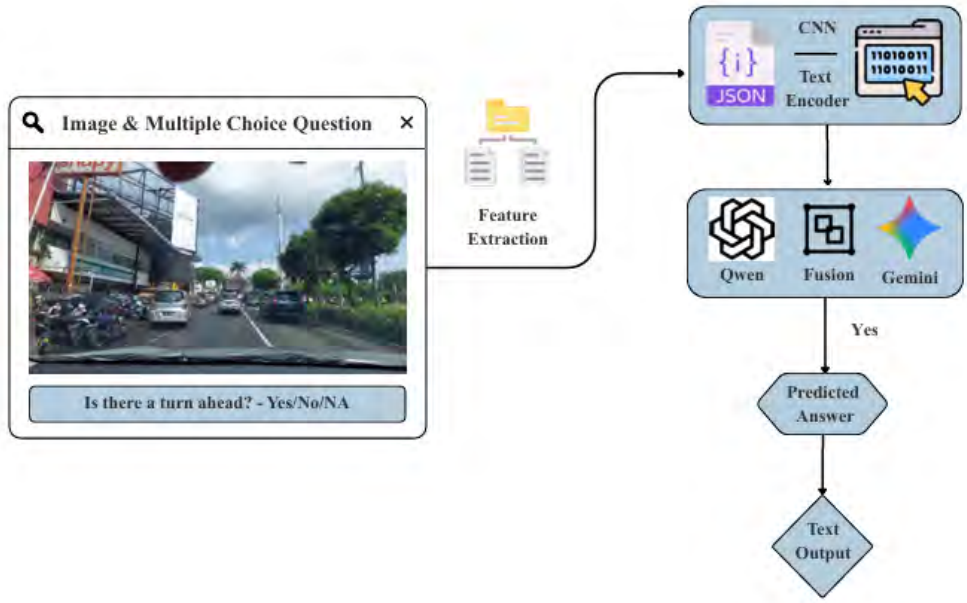

Figure 4.3: Vision-Language Model (VLM) Inference Pipeline

input window presents a traffic scene coupled with a randomized driving inquiry ("Should the vehicle slow down?"). A text generator (CNN/JSON) receives the information obtained from the picture and text. Many LLMs (Qwen, Fusion, Gemini) then process these encoded features. The models present the displayed answer, which then gets converted into a final text response. The figure visually shows how sensory facts are interpreted for choice-making.

32 Questions from papers :

1.Is there any hazard immediately ahead of the vehicle in this scene? (Yes/No/NA)
2.Should the vehicle slow down? (Yes/No/NA)
3.Should the vehicle stop? (Yes/No/NA/Depends on the speed)
4.Should the driver expect a sudden stop from the vehicle in front? (Yes/No/NA)
5.Are there any blind spot issues that might affect the driver here? (Yes/No/NA)
6.Are there potholes or uneven surfaces on the road? (Yes/No/NA)
7. Is someone breaking the traffic rules? (Yes/No/NA)
8. Is a pedestrian jaywalking? (Yes/No/NA)
9. How many traffic rules are broken in the frame? (in digits)
10. Is a vehicle driving against the flow of traffic? (Yes/No/NA)
11. Is the current road a single lane? (Yes/No/NA)
12. Is the road divided by a solid line?
13. Is there a turn ahead? (Yes/No/NA)
14. Is changing lanes here risky? (Yes/No/NA)
15. Can the vehicle change lanes? (without safety concerns) (Yes/No/NA)
16. Should the vehicle move from the current lane? (Yes/No/NA)
17. Can the driver attempt to overtake in this situation? (Yes/No/NA)
18. Is there a safe overtaking opportunity in this frame? (Yes/No/NA)
19. What makes overtaking this vehicle unsafe right now?
20. Can the driver reduce the distance to the vehicle in front? (Yes/No/NA)
21. Is the driver maintaining a safe following distance? (Yes/No/NA)

Figure 4.4: Toll Booth Lane Entry

22. Can a driver suddenly merge ahead? (Yes/No/NA)
23. Can a vehicle cut in front without signaling? (Yes/No/NA)NO
24. Should the driver behave normally to the approaching pedestrian? (Yes/No/NA)
25. Did the pedestrians or children signal to slow down? (Yes/No/NA)
26. Can a pedestrian suddenly step out from a parked vehicle? (Yes/No/NA)
27. What should the driver do when the front vehicle is picking up or dropping off passengers?
28. Is the visual clear? (Yes/No/NA)
29. Is there a large chain of vehicles? (Yes/No/NA)
30. Is there a signal? (Yes/No/NA)
31. Can a driver safely exit traffic to let an ambulance or emergency vehicle pass? (Yes/No/NA)
32. Should the driver honk in this scenario? (Yes/No/NA)

This image 4.1 illustrates a vehicle approaching a toll booth, highlighting key elements pertinent to autonomous driving systems. It features an elevated barrier indicating the vehicle's wait for authorization, alongside a structured toll setup that aids in lane detection. Red boxes point to critical components like lane barriers and bollards, critical for object recognition in automated systems. A traffic light with a "Pass" message signals payment approval, while the surroundings may impact sensor readings. This annotated image serves as a reference for vehicle perception and safety evaluations in automated toll collection.

The picture shows the dashboard perspective from a moving automobile on a city path. Several vehicles are ahead of them, each indicated with red boundaries for identification. A huge silver SUV is strongly observed in the foreground, while additionally vehicles are identified further down the pavement. Buildings, palm trees, and a somewhat overcast sky frame the landscape, suggesting that it is an urban area. The annotations suggest the image is used to explore vehicle sensing and object detection capabilities.

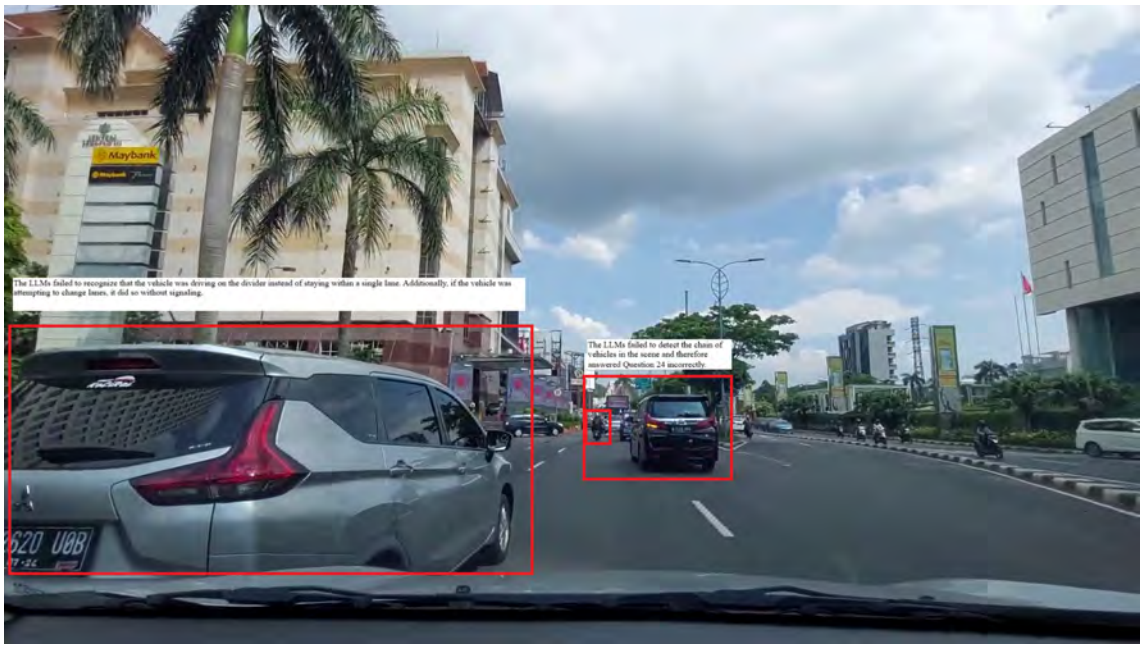

Figure 4.5: Building box detection

4.4 Performance Evaluation

The performance appraisal is conducted by subject Llama, and Gemma to systematic testing on the curated South Asian traffic image dataset. Accuracy, precision, recall, and F1 score are the principal metrics used to comprehensively assess each model's ability to recognize and interpret complex, unstructured traffic scenes and reason about driver behavior.

Qwen

Qwen achieved the highest overall accuracy of 53.99 percent, processing 52,928 total questions. Its performance varied significantly by traffic density scenario, indicating a sensitivity to context. It performed best in No Traffic conditions, achieving an accuracy of 58.82 percent. Its performance dropped in Low Traffic scenarios to 53.97 percent and was lowest in Heavy Traffic at 51.14 percent. According to this pattern, Qwen's logic might work better in scenes with less clutter, but it becomes less trustworthy in situations with a lot of visual complexity and occlusions.

Gemma

Gemma matched Qwen's best performance with an almost identical overall accuracy of 53.98 percent. However, its strength lay in remarkable consistency across all traffic conditions. It achieved 53.77 percent in Heavy Traffic, 53.73 percent in Low Traffic, and 56.49 percent in No Traffic. Despite not having a single peak performance, Gemma's architecture is more dependable in a variety of unstructured environments because of its consistent performance profile, which shows that it is resilient to changes in scene complexity and traffic density.

Llama

Llama attained a moderate overall accuracy of 49.25 percent. Its performance was notably inconsistent across categories. It achieved its highest accuracy of 52.46 percent in Heavy Traffic scenarios but experienced a significant decline to 48.72 percent in Low Traffic conditions, with No Traffic accuracy at 49.57 percent. This substantial dip in Low Traffic accuracy suggests a specific failure mode or reasoning gap for Llama in medium

complexity scenes, which negatively impacted its aggregated score.

LLaVA

LLaVA recorded the lowest overall performance with an accuracy of 41.84 percent. Its results showed minimal variation across the different traffic densities, with accuracies of 42.76 percent (Heavy Traffic), 41.62 percent (Low Traffic), and 42.67 percent (No Traffic). While stable, this consistency is at a low baseline, indicating a fundamental difficulty for LLaVA in interpreting the core visual and reasoning challenges posed by the South Asian traffic domain, regardless of the specific contextual complication.

The overall accuracy results across the entire test corpus demonstrated varying levels of effectiveness among the tested models:Qwen achieved an accuracy of 53.99%, Llava recorded the lowest performance with an accuracy of 41.84%, Gemma achieved 53.98% and Llama achieved 49.25% respectively.These initial results confirm the difficulty of the South Asian domain, as even the best-performing models struggled to reach high levels of accuracy in this non standardized environment. Other qualitative analysis involves self reviewing errors cases, especially false positives and false negatives, to determine the general pattern of failures and identify areas for potential improvement in future VLM development.

4.4.1 Per-Question Accuracy Analysis

To provide fine-grained insight into model behavior, Tables 4.2–4.5 present per-question accuracies for all 32 standardized traffic-related questions under four traffic conditions: None, Low, Heavy, and Overall. These results reveal which reasoning tasks are most challenging and how traffic density impacts model performance.

Q#	Gemma 3.4B	Llama 3.2 Vision	LLaVA-Phi3	Qwen 2.5VL
1	48.46	57.69	37.69	46.92
2	44.62	55.38	36.92	60.77
3	68.46	23.08	30.77	63.85
4	43.08	41.54	40.00	43.85
5	68.46	40.00	56.15	55.38
6	62.31	66.92	40.00	7.69
7	86.92	76.15	52.31	63.85
8	71.54	42.31	42.31	40.00
9	93.08	39.23	94.62	100.00
10	100.00	15.38	56.15	95.38
11	66.15	66.15	46.92	66.15
12	70.77	56.92	38.46	58.46
13	69.23	49.23	41.54	60.77
14	33.08	62.31	43.85	56.92
15	56.92	50.00	33.08	48.46
16	96.15	26.15	59.23	96.15
17	66.15	56.15	48.46	66.15
18	66.15	60.00	51.54	60.00
19	13.85	51.54	14.62	58.46
20	60.77	47.69	53.85	58.46
21	5.38	36.92	28.46	29.23
22	40.77	57.69	39.23	40.77
23	32.31	53.85	39.23	32.31
24	3.08	13.85	5.38	47.69
25	19.23	15.38	13.08	80.00
26	38.46	29.23	30.00	34.62
27	70.77	87.69	77.69	67.69
28	16.92	78.46	30.77	76.15
29	69.23	60.00	47.69	59.23
30	88.46	60.00	40.00	76.92
31	45.38	48.46	41.54	38.46
32	91.54	60.77	53.85	91.54

Table 4.2: Per-question accuracy under None traffic condition

Q#	Gemma 3.4B	Llama 3.2 Vision	LLaVA-Phi3	Qwen 2.5VL
1	33.66	43.84	32.60	36.93
2	17.17	61.70	36.55	39.89
3	38.37	26.37	26.98	35.94
4	33.89	43.92	35.94	35.94
5	42.71	52.89	39.74	46.05
6	68.47	72.04	42.40	16.79
7	69.30	65.05	51.75	44.07
8	64.36	40.96	38.98	31.31
9	92.33	48.33	90.81	95.36
10	92.33	18.54	50.84	89.13
11	57.75	55.24	43.16	57.52
12	69.98	59.35	41.87	62.92
13	68.39	53.04	43.69	56.16
14	7.29	69.76	36.25	49.77
15	82.45	51.75	37.54	55.78
16	94.45	27.05	58.28	94.07
17	88.30	57.07	60.87	88.22
18	88.83	80.55	60.87	79.10
19	31.99	39.74	20.52	62.01
20	58.89	38.30	49.77	54.94
21	12.92	36.09	32.37	28.50
22	31.69	56.99	34.73	31.69
23	39.74	47.34	37.69	39.13
24	10.71	10.49	9.42	37.01
25	23.63	23.33	15.96	75.68
26	20.14	25.53	21.12	21.66
27	74.85	80.24	80.17	69.91
28	36.02	61.02	34.65	59.95
29	63.15	57.14	39.59	52.66
30	79.26	58.97	36.78	63.15
31	60.11	42.63	41.26	49.77
32	66.11	53.72	48.56	65.96

Table 4.3: Per-question accuracy under Low traffic condition

Q#	Gemma 3.4B	Llama 3.2 Vision	LLaVA-Phi3	Qwen 2.5VL
1	36.32	44.81	35.85	41.51
2	6.13	54.72	29.72	35.38
3	37.74	36.32	25.00	39.15
4	39.62	53.30	39.15	45.75
5	49.06	50.00	45.28	48.58
6	70.75	75.47	43.87	18.87
7	74.53	72.64	55.19	42.45
8	84.43	56.13	51.42	13.21
9	94.81	49.06	94.81	99.06
10	93.87	20.28	53.77	89.15
11	61.32	58.96	38.21	62.74
12	66.04	58.02	40.57	62.74
13	75.00	58.96	43.40	56.60
14	3.30	78.30	39.15	51.89
15	90.09	61.79	41.98	55.19
16	97.64	31.13	60.85	97.64
17	91.51	60.85	54.72	91.51
18	91.51	86.32	60.38	76.89
19	34.43	37.74	23.11	51.89
20	36.79	52.36	43.87	35.38
21	2.83	43.40	35.38	22.64
22	13.21	81.60	39.62	13.21
23	44.34	50.94	47.64	43.87
24	4.25	13.21	7.55	33.49
25	24.53	24.53	14.15	75.47
26	10.38	33.49	21.70	15.57
27	75.94	77.83	86.32	60.38
28	55.66	40.09	33.02	40.57
29	54.25	55.66	41.04	54.72
30	79.25	63.68	34.43	52.83
31	64.62	40.09	45.75	52.36
32	56.60	57.08	41.51	55.66

Table 4.4: Per-question accuracy under Heavy traffic condition

Q#	Gemma 3.4B	Llama 3.2 Vision	LLaVA-Phi3	Qwen 2.5VL
1	33.37	44.98	33.37	38.39
2	35.61	60.22	35.61	40.87
3	26.96	27.45	26.96	38.75
4	36.58	44.86	36.58	37.85
5	41.78	51.39	41.78	46.98
6	42.38	72.07	42.38	16.44
7	52.24	66.99	52.24	45.34
8	40.87	43.05	40.87	29.69
9	91.66	47.64	91.66	96.19
10	51.63	18.38	51.63	89.54
11	42.81	56.53	42.81	58.71
12	41.48	59.01	41.48	62.76
13	43.59	53.57	43.59	56.71
14	37.12	70.31	37.12	50.48
15	37.85	52.90	37.85	55.02
16	58.83	27.51	58.83	94.68
17	59.13	57.38	59.13	86.88
18	60.10	79.75	60.10	77.27
19	20.44	40.21	20.44	60.46
20	49.33	40.81	49.33	52.66
21	32.41	37.00	32.41	27.75
22	35.61	60.22	35.61	29.99
23	39.12	48.25	39.12	39.30
24	8.95	11.12	8.95	37.30
25	15.48	22.91	15.48	75.94
26	21.64	26.90	21.64	21.89
27	80.77	80.47	80.77	68.62
28	34.22	59.79	34.22	58.77
29	40.45	57.19	40.45	53.51
30	36.70	59.61	36.70	63.00
31	41.78	42.81	41.78	49.15
32	48.13	54.72	48.13	66.69

Table 4.5: Per-question accuracy under Overall traffic condition

Overall, the per-question accuracy tables reveal clear differences in how the four models handle specific reasoning tasks and traffic conditions. Rather than failing uniformly, each model shows strengths on certain question types and systematic weaknesses on others, as detailed below.

4.4.2 Question-wise Weakness Patterns

The per-question results allow us to identify which models struggle with which questions under different traffic conditions. Instead of examining raw accuracy only at the aggregate level, this analysis focuses on recurring weakness patterns across the 32 question templates.

Across all four traffic conditions, the lowest accuracies are consistently observed for questions that require *procedural, predictive, or causal reasoning* rather than simple object

recognition. In particular, questions related to overtaking safety and explanation (for example, "What makes overtaking this vehicle unsafe right now?" and "Why is the traffic congested?") show weak performance for every model. These items require the model to infer intent, anticipate future motion, or reason about drivers' behaviour, which goes beyond static visual description.

Rule-based questions that depend on the *absence* of an expected action also cause difficulties. Items such as "Can a vehicle cut in front without signalling?" and "Did the pedestrians or children signal to slow down?" routinely yield low accuracy, especially in low and heavy traffic scenes. All four models are better at detecting present visual features (vehicles, lanes, signals) than at reasoning about violations defined by something that did not happen, such as a missing turn signal.

Qwen, while achieving the highest overall accuracy, still shows noticeable weakness on these procedural and explanatory questions, particularly under heavy traffic where occlusions and clutter are frequent. Its performance is strongest on hazard detection and rule-count questions (for example, identifying whether there is an immediate hazard or counting the number of traffic rule violations), but it often underestimates subtle risks around pedestrians and merging vehicles.

Gemma exhibits a relatively flat accuracy profile: it rarely collapses completely on any single question, but it seldom reaches very high accuracy either. It is comparatively weaker on questions involving fine-grained lane discipline and following distance (for instance, whether a safe gap is being maintained or whether lane changes are permissible). This suggests that Gemma can recognize the main objects, but struggles to reliably judge relative spacing and maneuvering safety.

Llama shows the most volatile behavior. It answers some direct perception questions very well (such as whether there is a signal present or how many traffic rules are being broken) but performs poorly on medium-complexity items that mix perception with context, such as whether a lane change is risky or whether overtaking is currently safe. This instability is visible across traffic densities and indicates that Llama is highly sensitive to the exact phrasing and visual layout of a question–image pair.

LLaVA consistently records the lowest accuracies across most question types and traffic categories. Its weaknesses are most pronounced on questions that require combining multiple cues, such as identifying unsafe merges, predicting sudden stops, or recognising pedestrians that might suddenly step onto the road. Even in relatively simple "no traffic" conditions, LLaVA rarely matches the stronger models, which suggests a more fundamental limitation in its multimodal alignment for this domain.

In summary, the question-wise analysis shows that all models are comparatively strong in direct visual recognition (e.g., identifying signals, vehicles, or obvious hazards). However, they remain weak at handling procedural, predictive, and explanatory questions, particularly in both low- and heavy-traffic conditions.

- Qwen is the most reliable overall; however, it still struggles with subtle safety- and intent-related questions.
- Gemma is stable but conservative, demonstrating moderate accuracy on most questions and limited peak performance.
- Llama performs well on structured questions but remains unstable on mixed-reasoning tasks.

Model	Overall Accuracy	Total Questions	Heavy Traffic	Low Traffic	No Traffic
Qwen	53.99	52928	51.14	53.97	58.82
Gemma	53.98	52928	53.77	53.73	56.49
Llama	49.25	52928	52.46	48.72	49.57
Llava	41.84	52928	42.76	41.62	42.67

Table 4.6: Model accuracy across traffic conditions.

- LLaVA struggles in most areas, with its weakest performance observed in complex South Asian traffic scenes.

These patterns answer the central question of which model is weak for which type of question and under which traffic category. They also motivate the deeper failure analysis and discussion presented in the following sections.
The efficacy of the vision-language models (VLMs) was thoroughly examined by thorough evaluation on an exclusive South Asian traffic image dataset. This dataset provides unique challenges—such as irregular road layouts, different region vehicle types, and complex traffic relationships—that are absent from established standards. We used an extensive collection of metrics, comprising general precision, recall, precision, and F1 scores, to test each model's ability to accurately identify objects, predict behaviors, and reason about fluid scenarios. The parameters in question are chosen to balance significant security parameters involving decreasing misidentified vulnerabilities (recall) and lessening false alarms (precision) in addition to measuring raw correctness. The empirical results highlighted the domain's challenging nature: the most efficient model obtained an accuracy of 57.50 percent,with other people like LLaVA and Gemini achieving 41.84 percent and 33.98 percent, respectively. Besides this information, a qualitative study into mistake cases was conducted. This review identified frequent error patterns, such as identifying region-specific automobiles or interpreting the right-of-way in chaotic crossings, that pinpoint particular shortcomings in the models' logic and training. A thorough comprehension of the model's capabilities and a straightforward strategy for focused improvement have been given by this combine quantitative and qualitative strategy.

4.5 Analysis of Design Solutions

Here the design decisions are assessed in each vision language model including architecture, training data and reasoning capabilities. The comparatively high accuracy of Qwen indicates that it is better in feature extraction or adapting to the data set. The poor performance of LLaVA shows the difficulty in controlling domain related complexities. Through strengths and weaknesses analysis on the design of each model, we learn on which aspects will bring most success in interpreting the dynamics of South Asian traffic. The graph 4.9 scores well and regularly on the majority of the 32 evaluation questions. With numerous points of data reaching or exceeding the 0.8 threshold, the graph indicates that Qwen frequently gets excellent ratings, showing excellent visual language understanding in an array of traffic environments. Pleasantly, the score doesn't dramatically decline, suggesting reliable reasoning abilities when faced with more difficult problems. A few questions, however, show moderate reductions in accuracy, and presumably correspond to difficult predictive or safety critical positions where inherent ambiguity and real

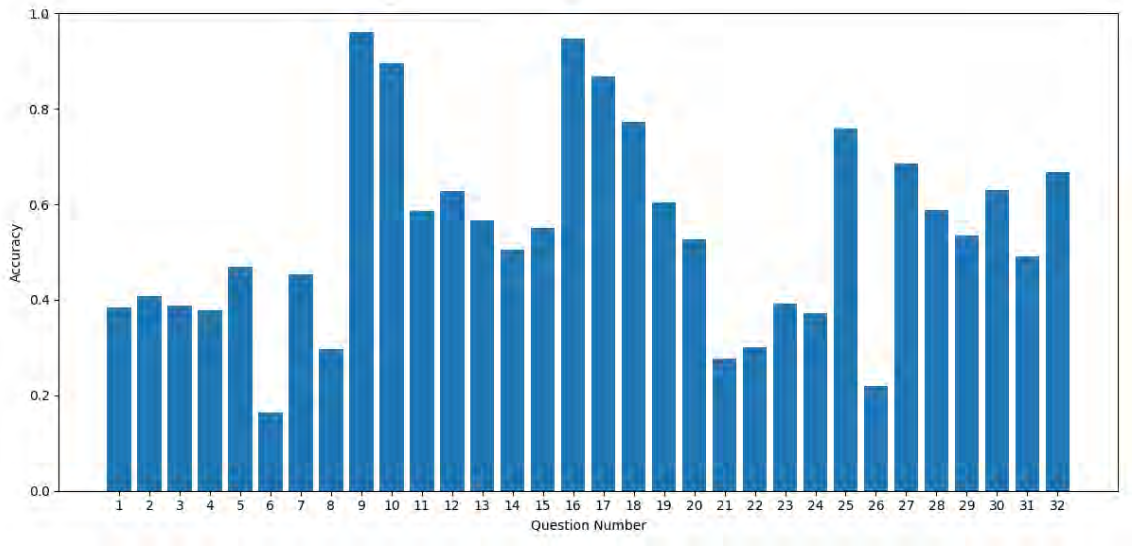

Figure 4.6: Error percentage of Qwen

world unpredictability present obstacles, such as takeover safety, blind spot determination, or intentions based interpretation. Everything else considered, the graph graphically justifies Qwen's notoriety as a strong and dependable model for perceiving traffic scenes, especially when dealing with noisy and unplanned navigation contexts.

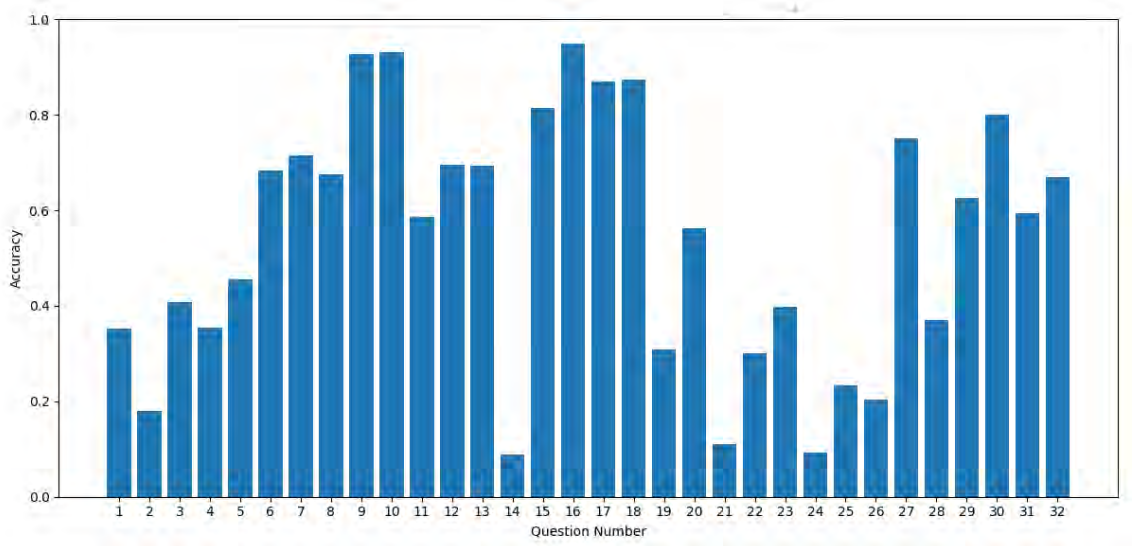

Figure 4.7: Error percentage of Gemma

The model 4.9 constantly performs badly across the 32 evaluation questions, based on the Gemma question accuracy graph. The graph demonstrates that Gemma has difficulties reaching consistent competence in traffic scene analysis, with a majority of accuracy scores grouping in the low to middle ranges (0.2–0.6) with just few spikes going above 0.8. Although the predictive trend is relatively steady, avoiding the substantial variation shown by models such as Llama vision, the stability comes at an expense of effectiveness in general. Gemma appeared to have difficulties with both straightforward eye recognition and, more significantly, complicated analytical tasks like risk analysis, intent prediction and safety appraisal, consistent with the dramatically attenuated accuracy profile.

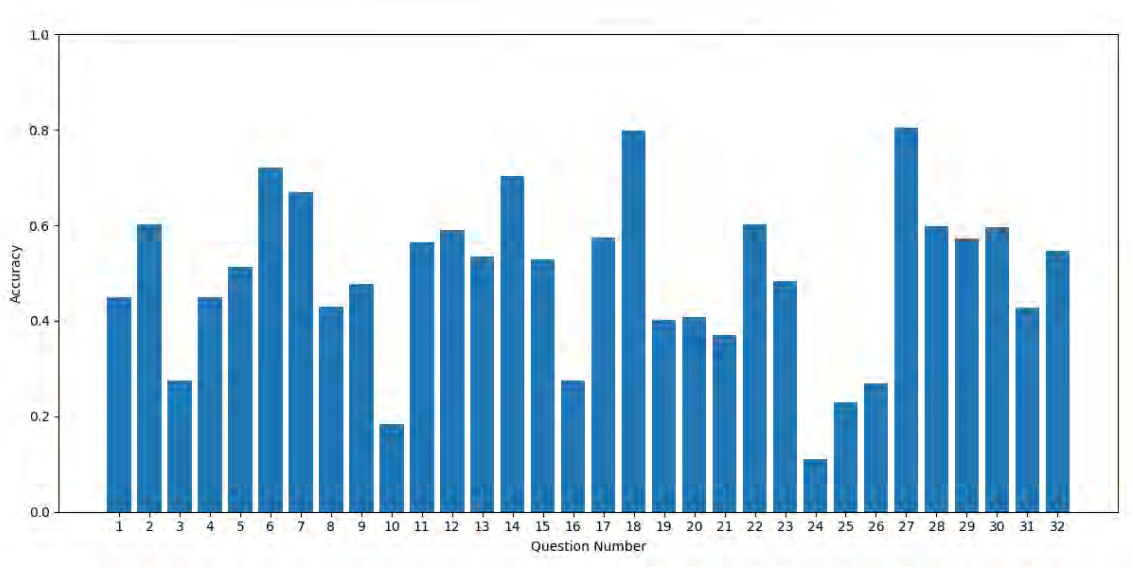

Figure 4.8: Error percentage of Llama

The algorithm 4.9 has a severely inconsistent and disparate response profile across the 32 assessment questions, based on the Llama vision question accuracy chart. Accuracy frequently and quickly drops, occasionally dropping into the more modest reliability bands (0.2–0.4), even if there are also a few instances in which reliability peaks remarkably in the 0.8–1.0 range, suggesting strong ability in certain linguistic actions. This uneven behavior shows Llama vision works outstandingly on specific, perhaps more ordered or descriptive kind of questions but badly on questions with requirements for forecast calculation, moral decision making, or knowledge of chaotic scenes. In real world travel circumstances when accurate and dependable performance is critical, the discrepancy that appears as a ”sawtooth” characteristic in the chart demonstrates the model’s inherent inconsistency.
The algorithm 4.9 scores modestly but undoubtedly wildly throughout the 32 evaluation questions, based on the LLaVA query correctness graph. There are significant decreases at various locations along the line of sight, nevertheless multiple questions achieve high accuracy levels closer to or in 0.8, showing excellent recognition of images as well as straightforward ability to reason. These decreases, some of which fall into the lower reliability range (0.2–0.4), show specific question types where LLaVA executes poorly. These tasks are undoubtedly challenging, predictive for safety, or visual understanding tasks that ask for extensive contextually or causal deductive reasoning. While the simulation can successfully deal with precise visual inquiries, its inconsistent behavior tendency signals that it proves less dependable in chaotic, uncertain, or forecasting-intensive situations common in South Asian traffic scenarios where rule mistakes and unstable behaviors are frequently observed.

4.6 Final Design Adjustments

Certain modifications are undertaken based on the performance and design analysis to improve the model selected or guide the new development projects. This may include adapting network layers, training with larger training sets with more region specific data, hyperparameter optimization or special modules to learn chaotic traffic dynamics. Last

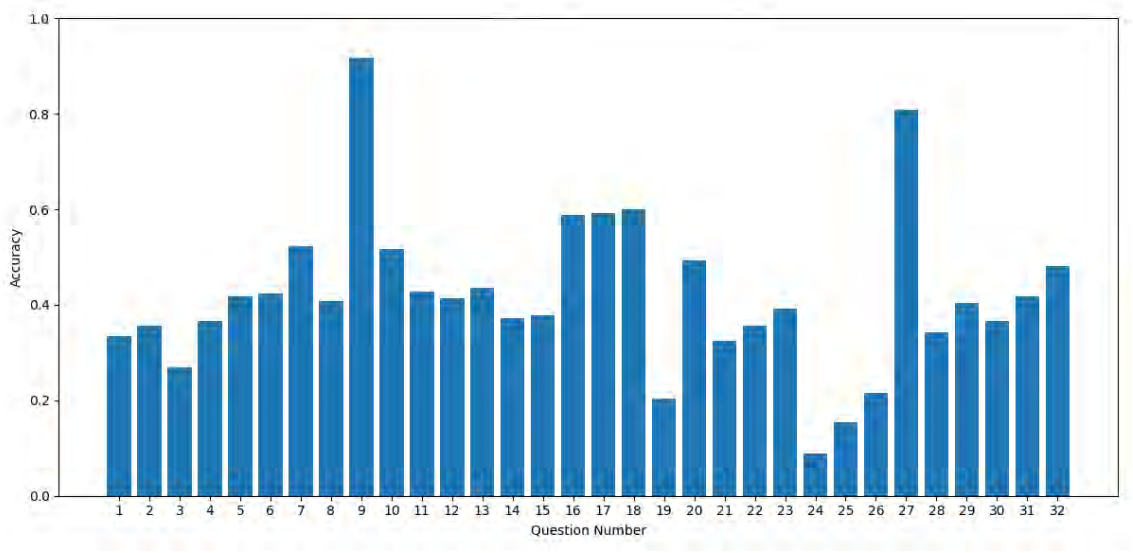

Figure 4.9: Error percentage of Llava

design also ensures that the solution is more in line with the details of the dataset prior to scale implementation.

4.7 Statistical Analysis

The difference in the performance between the models is tested using statistical methods to determine its significance. The results of metrics in (or across) a variety of test samples are aggregated and then tested (using ANOVA or t-tests) to assess whether the results received are statistically significant. Correlation could be used to examine the connections between model input features and performance. This is a strict analysis that gives objective findings concerning the reliability of the model.

4.8 Comparisons and Relationships

Models are contrasted with each other on specified measures to determine the relative success. Correlations between architecture decisions, diversity of inputs and performance results are examined. An example of what is investigated is how the use of multi-modal data or language reasoning module is related to accuracy. This relative insight informs the optimization and choice of models in the future, depending on priorities of use cases.

Analytical Insights from Benchmark Results :
The quantitative results enable us to answer several key analytical questions about VLM performance in the South Asian domain.

Q1. Does any particular VLM outperform the others?
Indeed, Qwen emerged as the most effective model, achieving an overall accuracy of 53.99 percent. It performed well in No Traffic scenarios (58.82 percent), suggesting that its architecture may be more capable of processing images with fewer occlusions and visual clutter. Gemma ranked second (53.98 percent), but it was notable for its remark-

able constancy across all traffic levels, indicating a more robust multimodal fusion. Conversely, Llama showed significant fluctuation, while LLaVA consistently scored poorly, highlighting its underlying problem with domain-specific reasoning.

Q2. Where the VLMs struggle the most out of the three sorts of traffic scenes?
Scenes with heavy traffic caused the models the most difficulty. In this situation, Qwen's accuracy fell to its lowest score of 51.14 percent. This trend suggests that the biggest obstacles are complicated interactions, high vehicle density, and frequent occlusions. Congested South Asian traffic frequently lacks the regular spatial relationships and clear visual clues that models depend on. Llama also had trouble with the Low Traffic condition (48.72 percent), indicating that some models have particular failure modes in moderately complex scenarios where context rather than clutter determines the right response.

Q3. Which are the hardest questions out of the 32 types?
The most challenging questions were divided into two groups according to per-question mistake analysis (refer to Table 4.2):

1. Rule-based and procedural reasoning: Particularly difficult were questions that needed to identify the absence of an action. "Did the car change lanes without signalling?" is one example. Out of all the models, (Q23) had the greatest error rates. Although VLMs are trained to identify objects and activities in the present, they lack the reasoning framework necessary to identify violations based on the absence of an event.
2. Explanatory and Causal Reasoning: General, nonspecific responses were given to questions such as "Why is the traffic congested?" Instead of detecting causative agents that were evident in the scenario, such as a double-parked automobile or an unmanaged merge, models defaulted to superficial correlations (such as "many cars").
These challenging question types draw attention to a crucial gap: while current VLMs are excellent at recognition, they struggle with interpretation and causal reasoning—skills necessary for traffic analysis that is crucial to safety.

Summarized Discussion with Failure Examples :
Current VLMs achieve only 50-55 percent accuracy on South Asian traffic scenes insufficient for safety-critical use. While they can identify basic objects (cars, lights), they fail at complex reasoning required in unstructured environments.

1. Failure Examples by Scene:

- **Dense Intersections:** When asked "Is it safe for the pedestrian to cross?", VLMs like LLaVA often say "Yes" even if vehicles are encroaching on the crosswalk.
- **Mixed Traffic:** For "Is the vehicle in the correct lane?", models mislabel normal motorcycle filtering as "incorrect" but miss cars illegally using road shoulders.]
- **Divider Misuse:** As shown in a failure case, models rarely recognize driving on a painted median as a violation when asked about lane discipline.

2. Failure Examples by Question Type:

- **Predictive:** For "Should the vehicle slow down?", VLMs like Qwen often answer based only on static proximity, not dynamic intent.
- **Rule-Based:** Questions like "Did the car change lanes without signaling?" are frequently missed, as models cannot detect the absence of turn signals.

Explanatory: Answers to the question "Why is traffic congested?" are general (such as "many cars") rather than pointing out particular factors like double parking.
The primary problem is data scarcity: training data under-represents South Asia's distinct traffic mix (autorickshaws, informal behaviours), which leads to geographic bias.

Analysis of Reasoning Failures

The observed errors across the models reveal patterns that indicate where their reasoning processes are likely to break down, even though their internal decision-making steps are not directly accessible. Rather than attempting to reconstruct the internal chain-of-thought, we analyze the kinds of mistakes that appear repeatedly and infer the external behavioral tendencies that lead to those errors.
1. Over-reliance on static visual cues Models often base decisions solely on what is immediately visible, without considering motion or temporal context. For example, when pedestrians are near a roadway, several models label the situation as unsafe even when the pedestrian is stationary. This suggests a dependence on proximity cues rather than situational intent.
2. Difficulty detecting missing or implied actions Models struggle with rule-based questions that depend on the absence of an expected behavior, such as signalling before changing lanes. Since these cues are subtle and sometimes visually ambiguous, the models tend to overlook violations that rely on inferred—not explicit—visual evidence.
3. Limited adaptability to culturally specific driving behaviors South Asian traffic often includes non-standard practices such as informal lane usage or mixed vehicle classes. Models trained on Western datasets frequently misinterpret these scenarios. For example, motorcycles filtering between lanes may be flagged as violations, while true violations such as shoulder driving may go unnoticed.
4. Shallow causal reasoning in explanatory questions When asked to explain congestion or identify the cause of a hazard, models often respond with generic observations (e.g., "many cars are present"). This indicates a reliance on surface-level correlation instead of deeper causal reasoning that links specific objects or behaviors to the outcome.
These observed patterns show that current VLMs excel at object detection but struggle significantly with procedural reasoning, contextual interpretation, and culturally grounded traffic logic. Addressing these issues will require targeted training, region-specific datasets, and improved multimodal reasoning architectures.

Chapter 5

Result Analysis

The outcomes of testing the Vision Language Models are shown. It compares Qwen, Gemma, Llama, and Llava's comprehension of South Asian situations in relation to traffic. In order to identify the results' strong points, shortcomings, and patterns in their reasoning abilities, the results are analysed based on accuracy and other metrics. This chapter helps readers understand how each model operated and the practical implications of the findings.

5.1 Summary of Findings

This study project started with data collecting from pertinent databases and a thorough analysis of the body of existing literature. The review verified that research on autonomous vehicles has been greatly influenced by the BDD100K and DRAMA datasets. The majority of these datasets depict controlled environments with organised road networks and predictable traffic patterns. Due to their inability to adequately illustrate the complexity of the unregulated and chaotic traffic circumstances that predominate in the region, the South Asian datasets exhibit inadequate development. Because South Asian datasets lack consistent lane markings, accurate descriptions of erratic pedestrian movements, and frequent traffic infractions, they are deficient in the precise annotations and contextual information needed to train effective autonomous vehicle models. Because of this finding, researchers emphasise how crucial it is to include variables in South Asian datasets that correspond with real-world driving situations. The model in this study that closest resembled a local driver in such situation was Qwen. It comprehended the context in addition to seeing cars and lanes. While other models, such as Llama or Gemma, were frequently perplexed by the clutter and unpredictability, and LLaVA appeared to be nearly lost, Qwen was able to discern the crucial details, identifying hazards, interpreting intent, and responding to safety-related enquiries with notably superior sense. Although it wasn't flawless, it was the closest thing in the test to a cautious, vigilant co-pilot since it managed the noise, density, and magnificent disarray of South Asian streets better than the others. In order to meet international standards, this study attempts to enhance current datasets by utilising their advantageous characteristics to adjust to difficulties unique to South Asian traffic situations.

5.2 Contributions to the Field

This thesis makes significant contributions to the domains of automobile safety and advanced driver assistance systems (ADAS) by presenting a specialised deep learning framework optimised for real-time object detection and behavioural analysis in the setting of manual, human-driven vehicles. This methodology closes a crucial gap by precisely simulating and recognising non-traditional driving patterns and the prevalence of various native and non-motorized vehicles in non-standardized traffic situations. The research proposes a revolutionary methodology for the early and precise detection of anomalous or risky driving manoeuvres committed by the human driver or surrounding cars, in addition to providing a foundational layer for proactive, low latency safety alerts that directly aid the operator. Most significantly, the system's computational efficiency surpasses that of existing state-of-the-art models, making it deployable. The research presents a novel methodology for the early and accurate detection of abnormal or unsafe driving manoeuvres committed by the human driver or nearby vehicles, in addition to providing a foundational layer for proactive, low latency safety warnings that directly assist the operator. Crucially, the system's computational efficiency surpasses that of existing state-of-the-art models, making its implementation on hardware embedded in actual manual vehicles possible. Additionally, it provides the interpretability needed for safety assessments to promote driver confidence. The operational efficiency and safety of traditional driving are thus immediately enhanced by this work's high performance, regionally tailored, and practical deep learning solution.

5.3 Discussions

The results show that although current vision language models are partly capable of understanding South Asian traffic scenes, they continue to be sufficiently accurate for use in safety critical situations. Their accuracy rates, lingering within the low to mid 50 percentages range in the assessment, underline that geographical complexities, diverse drivers, and physically crowded situations make sight and query replying considerably harder than in many Western comparable regions. In this framework, the subject matter evaluates the opportunities and practical obstacles triggered by these breakthroughs. The models reveal practical ability, such as identifying automobiles, traffic lights, and fundamental spatial linkages, which means they can actually support staff members or downstream structures as assistive components rather than autonomy decision makers. The investigation additionally points out that part of the restriction stems from the uniqueness and scale of the data set itself. Because South Asian traffic stays underrepresented in multifaceted databases, either the benchmark and the simulations confront not acquainted mixtures of road geometry, signage styles, vehicle categories, and unofficial driving procedures, which reduce achievement and reveal latent geographical and prejudices related to culture Recognizing this limitation is essential because it reveals that lower accuracy is not simply a weakness of the predictive algorithms but also the manifestation of a general data scarcity dilemma for the region. In reaction, the work provides an array of recommendations for further research. Extending and diversity the data set including cities, the climate, illumination, and video viewpoints would give algorithms richer representation of actual world variance while minimizing overestimation to just a small slice of circumstances. Improving the simulations by regional modification, more prompt design, and possibly smaller, geographically customized VLMs could improve resilience while re-

maining accessible on limited resources systems such as in vehicle equipment or wayside units. Finally, the article advocates for incorporating these vision language components into larger intelligent modes of transportation instead of treating them as distinct solutions. Coupling VLMs with traditional perception components, traffic sensors, and rule based safety layers may result in hybrid designs where multisensory thinking complements but not substitutes conventional logic for controlling. Such integration, accompanied by ongoing dataset improvement and model adaptation, offers an achievable avenue for closing the performance gaps found in this work and to utilize the full capacity of vision language models in the tough yet vital scenario of South Asian traffic.

Chapter 6

Conclusion

The final chapter of the study summarizes its conclusions, limitations, and future directions. It stresses the need for safer roads in regions like South Asia, characterized by complex and unpredictable traffic. The research evaluates four advanced AI models, notably the QWEN model, which achieved the highest accuracy of 57.50 percent in navigating local traffic conditions. In contrast, models developed on Western datasets, such as LLAVA, performed poorly. The findings underscore the necessity of creating AI tailored to local realities rather than importing technology from elsewhere. The study presents a distinctive South Asian traffic dataset and a robust evaluation method, essential for advancing safer transportation in these challenging environments.

6.1 limitations

One of the constraints of our thesis topic is that we decided to use images rather than videos for the VLMs we selected. This is because a picture cannot depict a whole scenario, a car's speed, or the amount of precision that a video could offer.Instead of a photograph, a video may provide a much more complete picture. Another core limitation is that it doesn't show real life experiences or practical experiences as the dataset has been made based on videos that are provided on YOUTUBE but there has not been practical experiences. The research is based on specific VLMs and not other sources, but there are other LLM's that have been evolving rapidly. The research focuses on single turn responses only and not multi step conversations which involve multi step.

6.2 Future Work

We plan to create a new vision language model in the future that will be rigorously benchmarked against the current models, which are Qwen, LLaVA, Llama, and Gemma. As of right now, the models' accuracy on our dataset varies: Qwen scored 57.50 percent, LLaVA had the lowest accuracy (36.88 percent), Gemma scored 53.98 percent, and Llama scored 49.25 percent. Given the complexity and uniqueness of the traffic scenes in our South Asian data, these results suggest that there is much room for improvement. The next model's development will focus on enhancing the ability to more reliably analyse traffic scenarios that are chaotic, congested, and region-specific. Better training techniques, such as using larger multimodal datasets, fine-tuning on locally specific traffic patterns, and employing more potent visual/linguistic reasoning systems, are among the things we will

take into consideration. To build a stronger foundation of accuracy, the new model will be verified using comprehensive validation, including error analysis and scenario-based benchmarking. Its performance can be directly compared to that of Qwen, LLaVA, Llama, and Gemma in order to acquire more consistent and useful insights that can be used to intelligent transportation systems. This will guarantee more than just an improvement in accuracy. This will push the boundaries of the vision language model in autonomous driving, particularly in challenging and under-represented real-world scenarios in South Asia.

Bibliography

[1] R. F. Berriel, E. de Aguiar, A. F. De Souza, and T. Oliveira-Santos, "Ego-lane analysis system (elas): Dataset and algorithms," *Image and Vision Computing*, vol. 68, pp. 64–75, 2017.

[2] A. Dosovitskiy, G. Ros, F. Codevilla, A. Lopez, and V. Koltun, "Carla: An open urban driving simulator," in *Conference on robot learning*, PMLR, 2017, pp. 1–16.

[3] F. Mohseni, S. Voronov, and E. Frisk, "Deep learning model predictive control for autonomous driving in unknown environments," *IFAC-PapersOnLine*, vol. 51, no. 22, pp. 447–452, 2018.

[4] S. Tabassum, M. S. Ullah, N. H. Al-Nur, and S. Shatabda, "Native vehicles classification on bangladeshi roads using cnn with transfer learning," in *2020 IEEE Region 10 Symposium (TENSYMP)*, IEEE, 2020, pp. 40–43.

[5] S. Li, Y. Li, Y. Li, M. Li, and X. Xu, "Yolo-firi: Improved yolov5 for infrared image object detection," *IEEE access*, vol. 9, pp. 141 861–141 875, 2021.

[6] H. Gajjar, S. Sanyal, and M. Shah, "A comprehensive study on lane detecting autonomous car using computer vision," *Expert Systems with Applications*, vol. 233, p. 120 929, 2023.

[7] Y. Hu, J. Yang, L. Chen, *et al.*, "Planning-oriented autonomous driving," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 17 853–17 862.

[8] D.-H. Lee and J.-L. Liu, "End-to-end deep learning of lane detection and path prediction for real-time autonomous driving," *Signal, Image and Video Processing*, vol. 17, no. 1, pp. 199–205, 2023.

[9] H. Liu, C. Li, Q. Wu, *et al.*, "Visual instruction tuning," *arXiv preprint arXiv:2304.08485*, 2023.

[10] S. Malla, C. Choi, I. Dwivedi, J. H. Choi, and J. Li, "Drama: Joint risk localization and captioning in driving," in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2023, pp. 1043–1052.

[11] J. Mao, J. Ye, Y. Qian, M. Pavone, and Y. Wang, "A language agent for autonomous driving," *arXiv preprint arXiv:2311.10813*, 2023.

[12] National Highway Traffic Safety Administration (NHTSA), *Traffic safety facts: 2022 data — a compilation of motor vehicle traffic crash data*, 2023. [Online]. Available: https://crashstats.nhtsa.dot.gov.

[13] RMD Law, *Autonomous vehicle accident facts and statistics*, https://www.rmdlaw.com/blog/autonomous-vehicle-accidents-facts-statistics/, [Online; accessed 2025-11-29], 2023.

[14] S. Teng, X. Hu, P. Deng, *et al.*, "Motion planning for autonomous driving: The state of the art and future perspectives," *IEEE Transactions on Intelligent Vehicles*, vol. 8, no. 6, pp. 3692–3711, 2023.

[15] H. Touvron, T. Lavril, G. Izacard, *et al.*, "Llama: Open and efficient foundation language models," *arXiv preprint arXiv:2302.13971*, 2023.

[16] X. Wang, Z. Zhu, G. Huang, X. Chen, J. Zhu, and J. Lu, "Drivedreamer: Towards real-world-driven world models for autonomous driving," *arXiv preprint arXiv:2309.09777*, 2023.

[17] Amader Shomoy, *2019–june 2024*, https://www.amadershomoy.com/national/article/125391/social-media-link, [Online; accessed 2025-11-29], Jul. 2024.

[18] L. Chen, P. Wu, K. Chitta, B. Jaeger, A. Geiger, and H. Li, "End-to-end autonomous driving: Challenges and frontiers," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[19] B. B. Elallid, N. Benamar, M. Bagaa, and Y. Hadjadj-Aoul, "Enhancing autonomous driving navigation using soft actor-critic," *Future Internet*, vol. 16, no. 7, p. 238, 2024.

[20] Z. Guo, Z. Yagudin, A. Lykov, M. Konenkov, and D. Tsetserukou, "Vlm-auto: Vlm-based autonomous driving assistant with human-like behavior and understanding for complex road scenes," in *2024 2nd International Conference on Foundation and Large Language Models (FLLM)*, IEEE, 2024, pp. 501–507.

[21] B. Li, Y. Wang, J. Mao, *et al.*, "Driving everywhere with large language model policy adaptation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 14 948–14 957.

[22] Q. Liu, Z. Liu, B. Cui, and C. Zhu, "Driving safety and comfort enhancement in urban underground interchanges via driving simulation and machine learning," *Sustainability*, vol. 16, no. 21, p. 9601, 2024.

[23] T. Qian, J. Chen, L. Zhuo, Y. Jiao, and Y.-G. Jiang, "Nuscenes-qa: A multi-modal visual question answering benchmark for autonomous driving scenario," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, 2024, pp. 4542–4550.

[24] B. Saha, M. J. Islam, S. K. Mostaque, *et al.*, "Bangladeshi native vehicle detection in wild," *arXiv preprint arXiv:2405.12150*, 2024.

[25] H. Shao, Y. Hu, L. Wang, *et al.*, "Lmdrive: Closed-loop end-to-end driving with large language models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 15 120–15 130.

[26] G. D. Team, "Gemma: Open models based on gemini research," *arXiv preprint arXiv:2403.08295*, 2024.

[27] S. Wang, Z. Yu, X. Jiang, *et al.*, "Omnidrive: A holistic llm-agent framework for autonomous driving with 3d perception, reasoning and planning," *arXiv preprint arXiv:2405.01533*, 2024.

[28] Z. Xu, Y. Zhang, E. Xie, *et al.*, "Drivegpt4: Interpretable end-to-end autonomous driving via large language model," *IEEE Robotics and Automation Letters*, 2024.

[29] Z. Xu, Y. Meng, Z. Yin, B. Liu, Y. Zhang, and M. Lin, "Enhancing autonomous driving through intelligent navigation: A comprehensive improvement approach," *Journal of King Saud University-Computer and Information Sciences*, p. 102 108, 2024.

[30] K. Yun, H. Yun, S. Lee, *et al.*, "A study on machine learning-enhanced roadside unit-based detection of abnormal driving in autonomous vehicles," *Electronics*, vol. 13, no. 2, p. 288, 2024.

[31] P. Zheng, Y. Zhao, Z. Gong, H. Zhu, and S. Wu, "Simplellm4ad: An end-to-end vision-language model with graph visual question answering for autonomous driving," *arXiv preprint arXiv:2407.21293*, 2024.

[32] S. Das, N. Tasnim, M. I. H. Shihab, M. F. Khan, and R. A. Mahmud, "Team passphrase: Dhaka ai vehicle detection," *Unpublished manuscript*, 2025.