

# EIGHT WEEKS REPORT

CAPTIONING ANOMALOUS OBJECTS AND EVENTS  
ON INDIAN ROADS USING LARGE LANGUAGE  
MODELS FOR IMPROVING AUTONOMOUS VISUAL  
PERCEPTION AND DRIVER SAFETY



INTERNATIONAL INSTITUTE OF  
INFORMATION TECHNOLOGY

---

H Y D E R A B A D

# Contents

---

<b>1</b>	<b>Introduction</b>	<b>1</b>
<b>2</b>	<b>Related Works</b>	<b>2</b>
<b>3</b>	<b>Experiments</b>	<b>4</b>
3.1	Data Processing . . . . .	4
3.2	Methodology . . . . .	6
3.3	Implimentation Details . . . . .	8
3.4	Comparitive studies . . . . .	10
<b>4</b>	<b>Results</b>	<b>11</b>
<b>5</b>	<b>Future Work</b>	<b>18</b>
	<b>Bibliography</b>	<b>18</b>

## CHAPTER 1

# Introduction

---

The evolution of autonomous vehicles is transforming transportation, yet navigating the complex and unpredictable nature of Indian roads presents significant challenges. Indian roads are characterized by diverse traffic patterns, frequent pedestrian movements, erratic driving behaviors, and various obstacles, which often make traditional anomaly detection methods inadequate. These methods typically struggle to handle the chaotic and dynamic conditions unique to such environments. Large Language Models (LLMs) can play a crucial role in overcoming these challenges by providing advanced capabilities for detecting and captioning a wide range of anomalies. This includes identifying and describing lane changes, aggressive driving maneuvers, rule violations, and overspeeding, which enhances the situational awareness of autonomous systems. By integrating LLMs into autonomous vehicle systems, the technology can better respond to unexpected situations in real-time, thereby improving safety and reducing the likelihood of accidents. By integrating LLMs, which can offer detailed post-event analysis and insights into traffic patterns and incident causes, autonomous vehicles can achieve a more robust understanding of their surroundings. This integration aims to bridge the gap between the controlled scenarios of datasets like CODA [5] and the real-world unpredictability of Indian roads, ultimately enhancing the safety and efficiency of autonomous vehicles in diverse and challenging environments.



Figure 1.1

## CHAPTER 2

# Related Works

---

**Road Anomaly and Corner Case Datasets.** Li et al. (2023) introduced the CODA [5] (Corner case Object Detection Annotations) dataset, a pioneering effort in creating a real-world road corner case dataset for benchmarking self-driving vehicles. CODA addresses the crucial need for comprehensive data on rare and challenging road scenarios, providing a valuable tool for assessing and improving the capabilities of autonomous vehicle systems in complex real-world conditions. Building upon CODA, Wang et al. (2024) developed CODA-LM [2], the first Large Vision-Language Model (LVLM) benchmark for evaluating LVLMs on road corner cases. This work serves as an automatic and systematic evaluation tool for LVLMs in the context of autonomous driving. The researchers also introduced CODA-VLM, a new driving LVLM comparable to GPT-4V on CODA-LM, which surpasses open-source alternatives in driving perception and suggestions. Expanding on these advancements, Chandra et al. (2024) introduced the METEOR [1] dataset, which offers a novel approach to capturing the complexities of dense, heterogeneous, and unstructured traffic scenarios specific to India. METEOR provides over 2 million labeled frames and 13 million annotated bounding boxes for 16 unique traffic agents, focusing on rare and complex instances such as traffic violations and atypical interactions. The dataset includes detailed annotations of diverse scenarios and is tagged with factors like weather, time of day, and road conditions. By evaluating state-of-the-art object detection and multi-agent behavior prediction methods on METEOR, the research presents a fine-grained analysis of the relationship between traffic environments and perception, thus advancing the field of autonomous driving through its comprehensive and detailed data. Varma et al. (2019) introduced the IDD [6] (Indian Driving Dataset), aiming to address the challenges of autonomous navigation in unconstrained environments. While similar in scale and annotation to datasets like KITTI and Cityscapes, IDD significantly expands the scope by focusing on the complexities of Indian road conditions. The dataset comprises over 10,004 labeled images with fine instance-level boundaries, capturing a higher diversity of object classes and within-class variations. IDD includes novel traffic agents such as autorickshaws and animals, and accounts for a broad range of environmental factors including air quality and dust. Performance evaluations reveal significant performance drops in semantic segmentation models on IDD compared to Cityscapes, highlighting the need for more complex datasets and larger-scale training data. The unique characteristics of IDD, such as ambiguous drivable areas and diverse traffic participants, suggest new directions for research in domain adaptation and behavior prediction, marking a substantial contribution to advancing autonomous navigation in diverse and less structured environments.

**Anomaly Detection.** Li et al. (2023) used the CODA dataset to evaluate state-of-the-art object detectors such as Cascade R-CNN, Deformable DETR, and Sparse

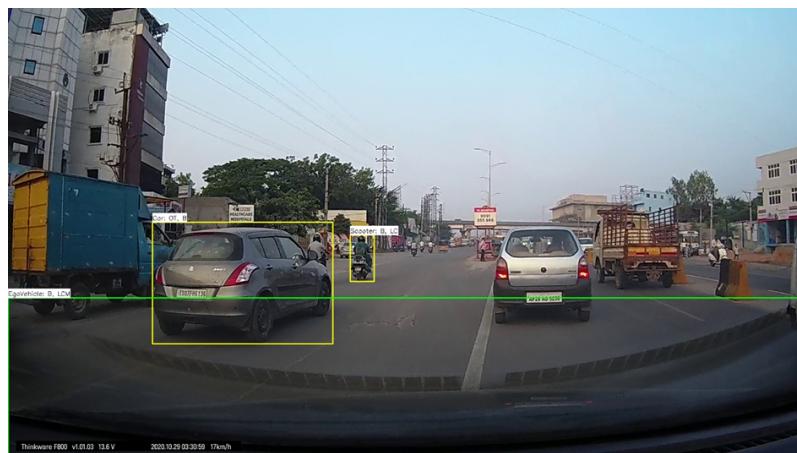
R-CNN. Their work highlighted the significant challenges these systems face in handling complex and rare scenarios, underscoring the need for more robust detection mechanisms in autonomous vehicles. Wang et al. (2024) demonstrated that text-only Large Language Models (LLMs) can effectively judge LVLM performance in the context of road corner cases, showing stronger consistency with human judgments compared to LVLM judges. This finding opens new avenues for evaluating and improving object detection systems in autonomous driving. Zhang et al. (2024) introduced InternLM-XComposer2 [4], an advanced vision-language model excelling in understanding and free-form text-image composition. Leveraging Partial LoRA and a diverse data foundation, it outperforms existing open-source models and even surpasses some closed-source APIs in several benchmarks. The model demonstrates exceptional capabilities in processing complex multimodal data and detecting various anomalies. InternLM-XComposer2 represents a significant advancement in multimodal AI, with promising applications in autonomous driving and other real-world challenges. Chen et al. (2024) presented InternVL [3], a groundbreaking large-scale vision-language foundation model that aligns a large-scale vision encoder with LLMs for the first time. This novel approach enables InternVL to achieve unprecedented performance across a wide spectrum of generic visual-linguistic tasks, including visual perception, vision-language tasks, and multi-modal dialogue. The model introduces a innovative progressive image-text alignment strategy for efficient training, which strategically maximizes the utilization of web-scale noisy image-text data for contrastive learning and leverages fine-grained, high-quality data for generative learning. This dual approach allows InternVL to harness the vast amount of available data while maintaining high-quality outputs. InternVL demonstrates leading performance on various critical tasks such as image classification (ImageNet), semantic segmentation (ADE20K), video classification (Kinetics), image-text retrieval (Flickr30K COCO), video-text retrieval (MSR-VTT), and image captioning (COCO, Flickr30K, NoCaps). Notably, it also proves highly effective for multi-modal dialogue tasks (MME, POPE, Tiny LVLM), showcasing its versatility and potential for complex, real-world applications. The exceptional performance of InternVL across this diverse range of tasks positions it as a powerful tool for advancing autonomous driving technologies, particularly in areas requiring sophisticated visual understanding and multi-modal reasoning. These advancements in datasets and object detection models have greatly contributed to the development of more reliable autonomous driving systems. However, the unique complexities of Indian roads, characterized by unpredictable traffic patterns, diverse vehicle types, and erratic driving behaviors, necessitate further research to adapt and extend these models for even more challenging urban environments.

## CHAPTER 3

# Experiments

## 3.1 Data Processing

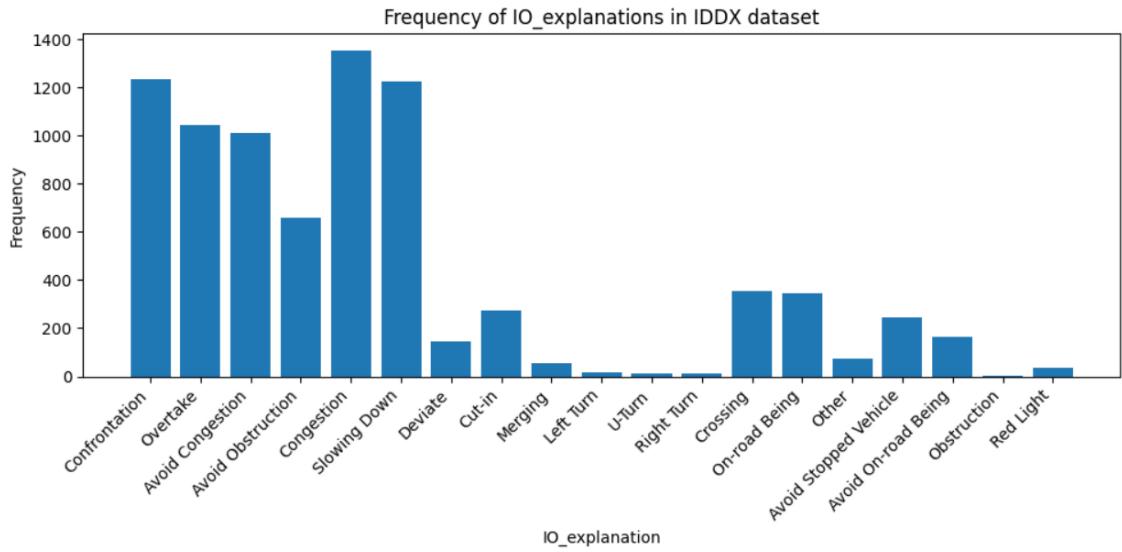
Video files from the METEOR [1] dataset were processed to extract and annotate frames with objects exhibiting specified "Out-Of-Distribution" (OOD) actions, such as 'LaneChanging(m)' (LCM), 'Cutting' (C), 'Yield' (Y), 'Behaviour' (B), 'OverTaking' (OT), 'LaneChanging' (LC), 'ZigzagMovement' (ZM), 'RuleBreak' (RB), and 'OverSpeeding' (OS). Frames were uniformly extracted, and XML files were parsed to retrieve annotations, including object IDs, names, bounding box coordinates, and attribute values for the OOD actions. Each frame was then checked for the presence of any OOD actions. If detected, bounding boxes were drawn around the relevant objects, labeled with their names and true attributes. The annotated frames were saved with filenames structured as frame-framenumber-objectids-attributevalues.PNG, using abbreviations for the attributes. For example, 'OverTaking' was abbreviated as 'OT'. Additionally, a list of frames that were not extracted due to the absence of specified OOD actions was maintained, and all unique object IDs encountered were printed. The METEOR dataset was chosen for its comprehensive representation of Indian traffic conditions, capturing diverse static anomalies (e.g., potholes, barriers) and dynamic anomalies (e.g., abrupt lane changes, overtaking, overspeeding).



**Figure 3.1:** Annotated traffic scene highlighting vehicles with bounding boxes, used for analyzing and detecting various traffic behaviors.

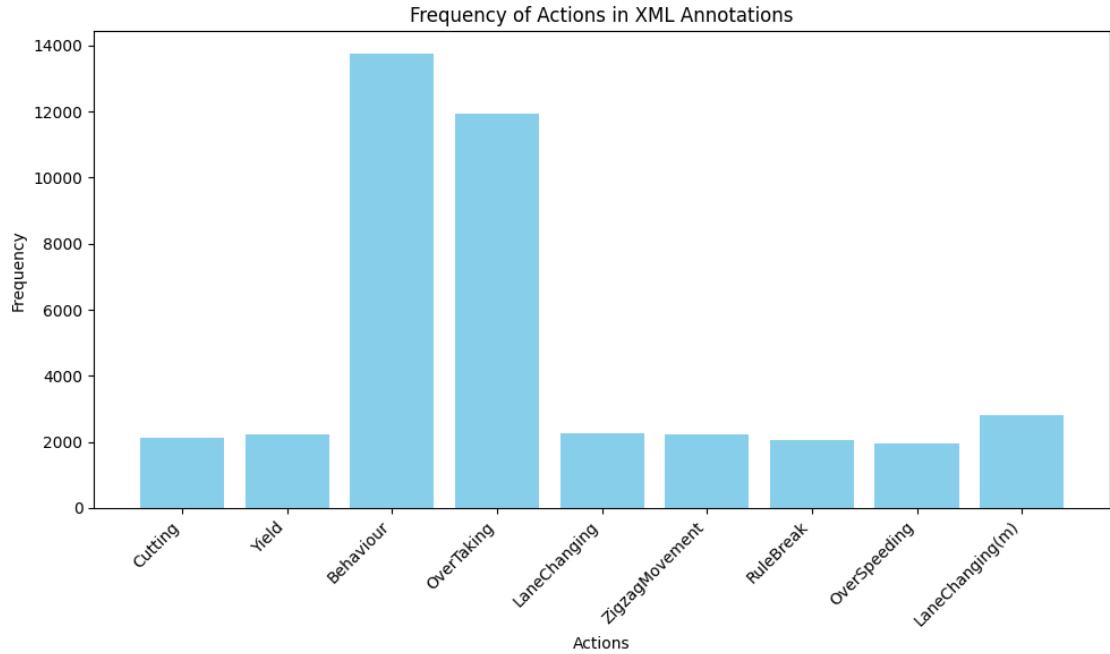


**Figure 3.2:** An example of annotation with anomalies marked in grey



**Figure 3.3:** IDD Statistics

**Further Analysis of IDD and METEOR Datasets** To gain deeper insights into the traffic behaviors captured in the IDD and METEOR datasets, comprehensive analyses were conducted for each dataset, focusing on the frequency and distribution of specific actions and scenarios. For the IDD dataset, a detailed analysis was performed to extract and count occurrences of various IOexplanation values from the dataset's JSON annotations. This revealed the distribution of different types of events, providing an overview of the dataset's balance. The results were visualized with a histogram to assess the prevalence of each action, aiding in understanding the dataset's diverse driving scenarios and its balance across different types of events. Similarly, for the METEOR dataset, XML annotations were processed to count specific traffic actions, such as traffic violations and dynamic behaviors. The total number of frames processed and the frequency of each action were reported. A histogram was generated to illustrate the distribution of these actions across the dataset, highlighting the dataset's focus on capturing complex and rare traffic scenarios. This analysis is crucial for evaluating the dataset's effectiveness in representing real-world traffic conditions and supporting the development and testing of autonomous driving systems.

**Figure 3.4:** Meteor Statistics

## 3.2 Methodology

**InternLM-XComposer2** The InternLM-XComposer2 model represents an advanced approach to multimodal learning by seamlessly integrating a vision encoder with a large language model (LLM) through a novel Partial LoRA module. At its core, the vision encoder, built on the OpenAI ViT-Large model, is tasked with extracting high-level visual features from raw images. This encoder, pretrained using image-language contrastive learning techniques, effectively transforms visual data into a form that can be leveraged for further processing. The resulting visual tokens are then passed to the Partial LoRA module, which specializes in aligning these tokens with textual information from the LLM. The Partial LoRA (Low-Rank Adaptation) module is a key innovation in this architecture. It adapts the traditional LoRA approach by applying low-rank adaptation specifically to the visual tokens, rather than treating all tokens uniformly. This selective adaptation allows the model to integrate new modality-specific knowledge without overwhelming the existing LLM capabilities. In practice, the visual tokens are enriched with new information while preserving the language tokens' inherent properties. The LLM used in this setup has demonstrated robust performance across various benchmarks. By combining the visual tokens with the tokenized text, the model creates a unified input sequence that leverages both visual and linguistic information. This approach not only enhances the model's ability to understand and generate responses based on multimodal inputs but also optimizes computational efficiency by minimizing the alignment costs traditionally associated with integrating different modalities. In summary, InternLM-XComposer2's architecture effectively merges high-level visual feature extraction with advanced language processing, enabled by the Partial LoRA

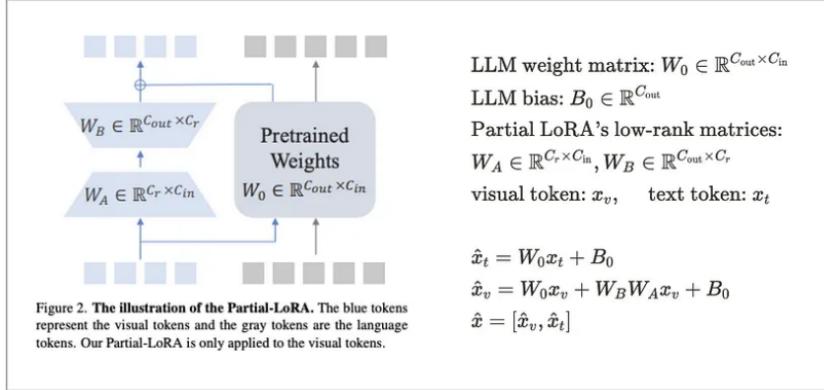


Figure 3.5: Image from original paper

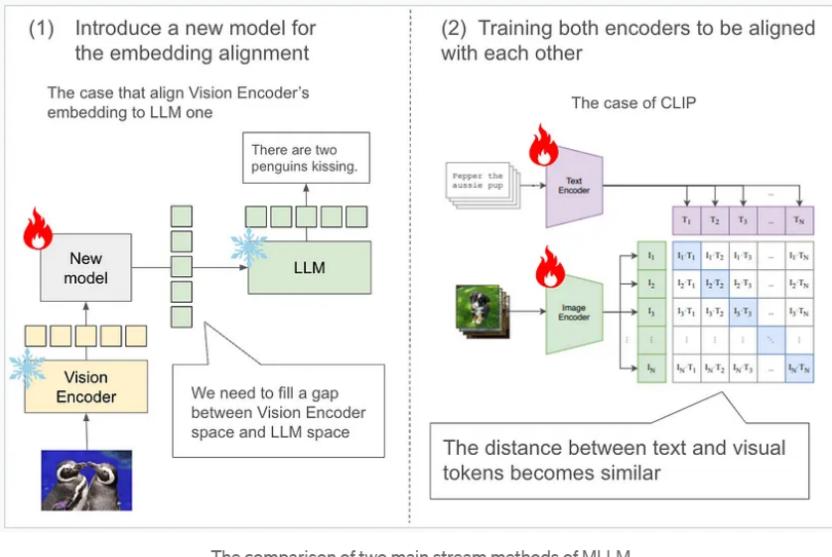
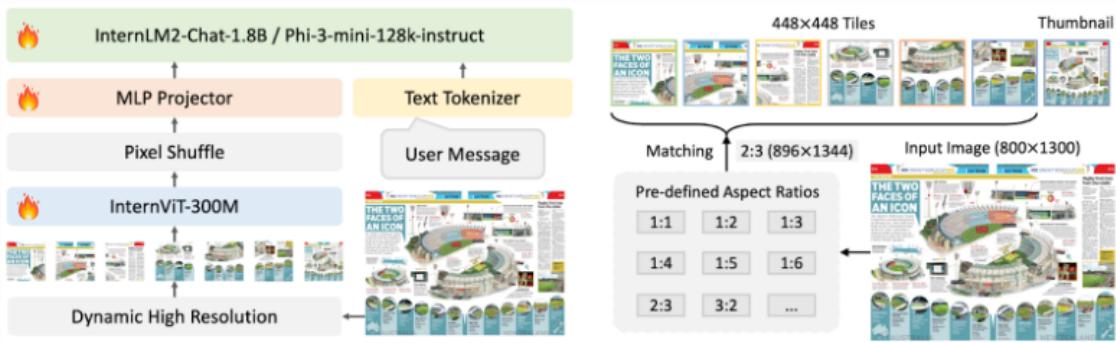


Figure 3.6

module's targeted adaptation. This integration supports sophisticated multimodal applications, offering significant improvements in understanding and generating contextually relevant responses.

**InternVL** InternVL is a sophisticated multimodal model designed to synergistically process both visual and textual information. At its core, InternVL comprises two primary components: the InternViT image encoder and the InternLM language model. The InternViT is adept at extracting meaningful visual features from input images by dividing them into 448x448 pixel tiles. This approach, coupled with the model's capacity for dynamic high-resolution handling, allows for robust processing of images with varying dimensions. Complementarily, the InternLM excels in comprehending and generating text, converting human language into a format suitable for machine processing. To bridge the gap between the visual and textual domains, InternVL incorporates an MLP projector. This component plays a pivotal role in mapping both visual and textual features into a shared embedding space, facilitating

**Figure 3.7:** InternVL Architecture

cross-modal interactions. By aligning these distinct representations within a unified space, the model can establish complex correlations between images and text, enabling tasks such as image captioning, visual question answering, and image-text retrieval. Furthermore, a pixel shuffle operation is employed as a preprocessing step to adjust image dimensions prior to feeding them into the InternViT. This ensures optimal input for the image encoder.

### 3.3 Implementation Details

**InternLM-XComposer2** The InternLM-XComposer2 model, a Large Language Model (LLM) optimized for multi-GPU settings, was employed to process the METEOR and IDD dataset, which focuses on detecting anomalies on Indian roads characterized by unstructured environments and chaotic traffic. The model was deployed on a system configured with four GPUs and four CPUs, each with 20 GB of memory. The architecture of the InternLM-XComposer2 model comprises multiple transformer layers designed to handle complex natural language processing tasks. The inputs to the model were images extracted from video frames and carefully designed text prompts. The outputs were captions describing the detected anomalies. To evaluate the model’s performance, the functionality was tested by inputting the designed prompts and video frames into the model. The output was analyzed for the presence of both static and dynamic anomalies. The processing time for each frame was recorded to assess the model’s real-time efficiency. This setup enabled the systematic identification and annotation of anomalous behaviors, thereby enhancing the analysis of traffic scenarios and contributing to improvements in autonomous visual perception and driver safety.

**InternVL** InternVL, the vision-language model under evaluation, was tested on an Ada server configured with 1 GPU and 36 CPUs, each equipped with 2GB of memory. The process began with the extraction of frames from video files of METEOR and IDD dataset at a fixed interval. These frames were captured using OpenCV and converted to RGB format for consistency in processing. Each frame was then transformed into a PIL image, resized, and split into blocks to accommodate the model’s input requirements. This step involved resizing the image to maintain an

appropriate aspect ratio and subsequently splitting it into manageable blocks based on the image size. To ensure compatibility with the model, frames underwent normalization using ImageNet mean and standard deviation values. This normalization was part of a comprehensive transformation pipeline, which included resizing and converting images to tensors suitable for model input. The pre-trained model, "OpenGVLab/Mini-InternVL-Chat-2B-V1-5," was loaded with 'torch.bfloat16' precision to optimize memory usage, and the corresponding tokenizer from Hugging Face's 'AutoTokenizer' was used for processing text inputs. For each frame, the model generated responses to specific prompts using its batch chat function. Multiple prompts were tried, and prompt engineering was done to obtain the "Yes" or "No" formatted captions for indicating the presence of obstacles. The responses were aggregated and analyzed to assess the model's performance. Time tracking was meticulously implemented to record both the total elapsed time for processing the video and the time taken for generating responses for each frame. This time tracking was crucial for evaluating the efficiency of the processing pipeline and the model's performance. Processed frames and corresponding input images were saved in a designated directory, with filenames reflecting frame-id and prompt details. In addition, text files were saved for the entire IDD dataset with filenames reflecting video names and prompt details, capturing responses and facilitating dataset creation for images with obstacles. Each text file recorded responses for each frame in a video, which were subsequently read and filtered. This filtration involved parsing the responses to determine whether obstacles were present in the frames. Specifically, the system checked the model's output for indications of obstacles, recording a "Yes" or "No" for each frame to denote the presence or absence of obstacles. These filtered results were then saved in a structured format, associating frame numbers with their respective obstacle indicators. Error handling mechanisms were incorporated to manage potential issues during frame processing, ensuring robustness and reliability of the results. This detailed setup not only provided insights into the model's efficiency and performance but also contributed to building a valuable dataset for obstacle detection. The saved images and recorded processing times offered a comprehensive view of the system's capability, laying the groundwork for future optimization and analysis.

## 3.4 Comparative studies

**InternLM-XComposer2** The model was tested using a variety of prompts to evaluate its performance in detecting anomalies in traffic scenes. The most effective prompts were selected based on their general applicability and success across most images. The model excelled in detecting static anomalies, identifying pit holes, barriers, concrete structures, potential accident-prone vehicles, pedestrians, and telephone poles. These detections are crucial for enhancing situational awareness in autonomous driving systems and preventing potential hazards. However, the model also generated some false positives, such as cones and construction signs as shown in fig 3.8, highlighting the need for further refinement to better distinguish between true anomalies and non-hazardous objects. While the model performed well with static anomalies, its efficiency in identifying dynamic anomalies was less impressive. The model struggled to accurately detect dynamic actions like abrupt lane changes, sudden stops, and erratic driving behaviors. This underscores the complexity of capturing the fluid and changing nature of dynamic traffic events.

**InternVL** The vision-language model InternVL was evaluated on the METEOR and IDD datasets. Multiple prompts were used during evaluation to determine the model's capability to detect obstacles. The most effective prompts were selected to generate "Yes" or "No" responses, which were then used to save text files indicating the presence or absence of obstacles in each frame. The model's performance on the IDD dataset was particularly noteworthy, producing satisfactory and coherent outputs that aligned well with the visual content. The evaluation process involved meticulous time tracking to record processing times, ensuring an efficient and robust assessment of the model's capabilities. InternVL demonstrated a strong ability to handle multimodal inputs and provided satisfactory answers for the IDD dataset, although there is room for improvement in detecting dynamic traffic anomalies. The overall results highlighted InternVL's potential for enhancing obstacle detection in autonomous driving systems, contributing to the development of more reliable and effective models.

## CHAPTER 4

# Results

---

InternVL and InternLM-XComposer2 showcased distinct strengths and areas for improvement in traffic anomaly detection. InternVL, tested on both METEOR and IDD datasets, proved effective in identifying static anomalies such as puddles and barriers, as highlighted by the example images. However, the model encountered challenges with pedestrian classification; specifically, when a person was present in the scene in Fig. 3.14, the model often failed to recognize it as an anomaly, inaccurately reporting "no obstacle" in such cases. This points to a need for refinement in distinguishing between true hazards and other scene elements. Conversely, InternLM-XComposer2 was evaluated using a variety of prompts to assess its performance. It excelled in detecting static anomalies like pit holes, barriers, concrete structures, and potential accident-prone vehicles, enhancing situational awareness for autonomous driving systems. Nevertheless, it produced some false positives, such as cones and construction signs, indicating the need for further refinement to differentiate between genuine anomalies and non-hazardous objects. The model also faced difficulties with dynamic anomalies, such as abrupt lane changes and erratic driving behaviors, reflecting the complexity of capturing the fluid and changing nature of traffic events. The confusion matrix provided further insights into InternVL's performance, revealing that while the model was highly accurate in confirming the absence of obstacles, it struggled significantly with detecting them. Specifically, the matrix showed that the model correctly identified 1,057 instances where obstacles were present and accurately recognized 3,341 instances where no obstacles were present. However, the model failed to detect 2,595 instances where obstacles were actually present, indicating a substantial gap in its detection capability. The results from both models are illustrated in the provided images, showcasing their successes and limitations. Additionally, text files were saved for the entire IDD dataset, documenting the model responses for each frame to indicate the presence of obstacles. This is crucial for further analysis and dataset enhancement, highlighting the ongoing need for model development to improve accuracy and handling of both static and dynamic anomalies.

		Actual	
		Obs.	No Obs.
Obs.	Obs.	1057	0
	No Obs.	2595	3341

**Figure 4.1:** Confusion Matrix for IDD Segmentation dataset using InternVL model.



**Figure 4.2:** (InternLM Model)Query:Are there any obstacles on the area outside the road? Describe the obstacle if any.

Output: Yes, there is a large concrete structure or pillar on the area outside the road. It appears to be part of an overpass or bridge construction.



**Figure 4.3:** (InternLM Model)Query:Are there any obstacles on the area outside the road? Describe the obstacle if any.

Output:Yes, there are several traffic cones placed on the area outside the road.



**Figure 4.4:** (InternLM Model)Query: Detect and describe any vehicles that are changing lanes or overtaking. Provide the type, color, and approximate distance of the vehicle from the camera. Output:In the image, there are two cars that appear to be changing lanes or overtaking. The first car is a silver hatchback, and it is located approximately 10 meters away from the camera. The second car is white and is positioned about 20 meters away from the camera. Both vehicles seem to be in motion, navigating through the busy street.



**Figure 4.5:** (InternLM Model)Query:Are there any obstacles on the area outside the road? Describe the obstacle if any.  
Output:There is a trash bag on the side of the road, which can be considered an obstacle.



**Figure 4.6:** (InternLM Model)Query:Are there any obstacles on the area outside the road? Describe the obstacle if any.

Output:Yes, there are several obstacles on the area outside the road. There is a large concrete structure or pillar that appears to be part of an overpass or bridge construction. Additionally, there are multiple barriers and fences set up along the side of the road, likely for safety purposes during the construction process.



**Figure 4.7:** (InternVL model)Query:Are there any obstacles or debris or barrier on the road ahead. Limit the answer to one sentence. Output:yes



**Figure 4.8:** (InternVL model)Query:Are there any obstacles or debris or barrier on the road ahead. Limit the answer to one sentence. Output:yes



**Figure 4.9:** (InternVL model)Query:Are there any obstacles or debris or barrier on the road ahead. Limit the answer to one sentence. Output:no



**Figure 4.10:** (InternVL model)Query:Are there any obstacles or debris or barrier on the road ahead. Limit the answer to one sentence. Output:yes



**Figure 4.11:** (InternVL model)Query:Are there any obstacles or debris or barrier on the road ahead. Limit the answer to one sentence. Output:no



**Figure 4.12:** (InternVL model)Query:Are there any obstacles or debris or barrier on the road ahead. Limit the answer to one sentence. Output:Yes, there is a pile of debris or barrier on the road ahead.



**Figure 4.13:** (InternVL model)Query:Are there any obstacles or debris or barrier on the road ahead. Limit the answer to one sentence. Output:Yes, there is a large puddle and there are barriers on the road ahead.

## CHAPTER 5

# Future Work

---

Future research will focus on enhancing the models' capabilities in identifying both static and dynamic anomalies in traffic scenes. Given the satisfactory performance of the models on static anomalies in the IDD dataset, efforts will be directed towards refining its detection accuracy for dynamic actions. This involves improving the model's sensitivity towards erratic driving behaviors, sudden stops, and abrupt lane changes. Another avenue of exploration is the development of more sophisticated prompt engineering techniques. By experimenting with a broader range of prompts and further refining the input queries, we aim to enhance the models' ability to provide precise yes/no responses for the presence of obstacles, thereby improving the quality of the annotated datasets. This will include expanding the prompt library and employing advanced natural language processing methods to better capture the nuances of traffic scene descriptions. Collaborative efforts will also be made to develop more robust evaluation metrics and benchmarking protocols, specifically tailored to the unique challenges posed by unstructured and heterogeneous traffic environments. These metrics will not only assess the accuracy and efficiency of anomaly detection but also evaluate the models' real-world applicability and reliability. Finally, future work will extend the dataset with more diverse traffic scenarios from various geographical locations, ensuring that the models are trained and evaluated on a wide array of driving conditions and cultural contexts. This will contribute to the development of universally applicable autonomous driving solutions that are resilient and adaptable to global traffic environments.

# Bibliography

---

- [1] Chandra, R., Wang, X., Mahajan, M., Kala, R., Palugulla, R., Naidu, C., Jain, A., Manocha, D.: Meteor:a dense, heterogeneous, and unstructured traffic dataset with rare behaviors (2022), <https://arxiv.org/abs/2109.07648>
- [2] Chen, K., Li, Y., Zhang, W., Liu, Y., Li, P., Gao, R., Hong, L., Tian, M., Zhao, X., Li, Z., Yeung, D.Y., Lu, H., Jia, X.: Automated evaluation of large vision-language models on self-driving corner cases (2024), <https://arxiv.org/abs/2404.10595>
- [3] Chen, Z., Wu, J., Wang, W., Su, W., Chen, G., Xing, S., Zhong, M., Zhang, Q., Zhu, X., Lu, L., Li, B., Luo, P., Lu, T., Qiao, Y., Dai, J.: Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 24185–24198 (June 2024), [https://openaccess.thecvf.com/content/CVPR2024/papers/Chen\\_InternVL\\_Scaling\\_up\\_Vision\\_Foundation\\_Models\\_and\\_Aligning\\_for\\_Generic\\_CVPR\\_2024\\_paper.pdf](https://openaccess.thecvf.com/content/CVPR2024/papers/Chen_InternVL_Scaling_up_Vision_Foundation_Models_and_Aligning_for_Generic_CVPR_2024_paper.pdf)
- [4] Dong, X., Zhang, P., Zang, Y., Cao, Y., Wang, B., Ouyang, L., Wei, X., Zhang, S., Duan, H., Cao, M., Zhang, W., Li, Y., Yan, H., Gao, Y., Zhang, X., Li, W., Li, J., Chen, K., He, C., Zhang, X., Qiao, Y., Lin, D., Wang, J.: Internlm-xcomposer2: Mastering free-form text-image composition and comprehension in vision-language large model (2024), <https://arxiv.org/abs/2401.16420>
- [5] Li, K., Chen, K., Wang, H., Hong, L., Ye, C., Han, J., Chen, Y., Zhang, W., Xu, C., Yeung, D.Y., Liang, X., Li, Z., Xu, H.: Coda: A real-world road corner case dataset for object detection in autonomous driving (2022), <https://arxiv.org/abs/2203.07724>
- [6] Varma, G., Subramanian, A., Namboodiri, A., Chandraker, M., Jawahar, C.: Idd: A dataset for exploring problems of autonomous navigation in unconstrained environments. pp. 1743–1751 (01 2019). <https://doi.org/10.1109/WACV.2019.00190>