IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 2, April 2025
Impact Factor: 7.67

SAFENAVGPT: LLM and Transformer Driven Anomaly Detection for Real-Time Visual Navigation

Vaisnavi N M[1] **and Dr. M. Praneesh**[2]
UG Student, Department of Computer Science with Data Analytics[1]
Assistant Professor, Department of Computer Science with Data Analytics[2]
Sri Ramakrishna College of Arts & Science, Coimbatore, Tamil Nadu, India

Abstract: *This research enhances the "VISIONGPT" framework for safe and efficient visual navigation, targeting visually impaired individuals and autonomous systems. It incorporates transformer-based models and attention mechanisms to improve object detection, scene understanding, and real-time anomaly detection. The current system employs YOLO-World for open-vocabulary object detection, a rule-based anomaly detection module, and LLMs (GPT-3.5/4) for scene interpretation and voice-based hazard alerts. While effective, traditional CNN-based models struggle with capturing complex object relationships and adapting to dynamic environmental changes.*

To overcome these limitations, we propose an enhanced architecture featuring Vision Transformers (ViTs) and attention mechanisms. ViTs refine object recognition, especially for detecting small or distant hazards, and enable the system to predict object movements and provide anticipatory alerts. The self-attention mechanism of transformers allows the model to dynamically weigh the importance of different objects, minimizing false positives and ensuring critical hazards are prioritized.

Experimental evaluations demonstrate improved detection precision, reduced false alarms, and better latency management, making the system more practical for real-time applications. The enhanced system supports dynamic scene transitions, proactive hazard warnings, and user-personalized alert mechanisms, making it adaptable to urban, indoor, and unpredictable environments.

This work advances LLM-assisted visual navigation, contributing to AI-driven accessibility solutions. The proposed architecture is scalable and can be integrated into autonomous vehicles, robotic systems, and assistive technologies. It also pushes forward the field of vision-language models and multimodal AI, showcasing how transformer-based models can significantly enhance real-time navigation safety and accessibility.

Keywords: Vision Transformers (ViTs), Attention mechanisms, Anomaly detection, LLMs (GPT-3.5/4)

I. INTRODUCTION

1.Statement of the problem

Advancements in machine learning and mobile computing have significantly improved object detection and segmentation, enhancing visual navigation, especially in dynamic urban environments. Traditional models like YOLO are effective in real-time object identification but struggle with complex scenarios due to their reliance on predefined class labels. Recent developments in Multimodal Large Language Models (LLMs) have enabled better integration of vision-language understanding, making them ideal for real-time anomaly detection in visual navigation. However, most research has focused on general visual assistance, not safety-critical applications

This paper introduces VisionGPT, a framework that combines LLMs with open-world object detection models for zero-shot anomaly detection, improving navigation safety, particularly for visually impaired individuals. By integrating real-time scene understanding and dynamic scenario adaptation powered by LLMs, the system enhances visual navigation.

DOI: 10.48175/IJARSCT-24908

The study also explores the effects of different prompt engineering techniques on system performance, offering insights for future advancements in vision-language-based anomaly detection.

2. Goals

The objective of this project is to develop an advanced real-time anomaly detection and visual navigation system by integrating transformer-based models, attention mechanisms, and large language models (LLMs). This research builds upon the existing VISIONGPT framework, which utilizes YOLO-World for object detection, a rule-based anomaly detection module, and GPT-3.5/4 for scene interpretation and voice-based hazard alerts. While effective, traditional convolutional models struggle with contextual scene understanding, dynamic environment adaptation, and predictive safety measures. To address these limitations, we introduce transformer-based enhancements to improve the system's overall accuracy, efficiency, and adaptability.

II. SOFTWARE SPECIFICATIONS

YOLO-World

YOLO-World is an open-vocabulary object detection model that can identify objects in real-time without relying on predefined class labels. It excels in dynamic environments, making it highly effective for detecting a wide range of objects in diverse scenarios.

LLMs

LLMs, like GPT-3 and GPT-4, are advanced models designed to process and understand multimodal data, combining text, images, and other forms of input. They are highly versatile, excelling in tasks such as natural language understanding, text generation, and real-time anomaly detection.

ChatGPT-3.5

ChatGPT-3.5 is an AI model developed by OpenAI, capable of understanding and generating human-like text based on a given prompt. It has applications in natural language processing tasks such as conversational agents, content generation, and problem-solving.

Vision

Vision Transformers (ViTs) are a deep learning architecture that applies transformer models to image data, offering a more efficient way of processing visual information compared to traditional convolutional neural networks (CNNs). ViTs are particularly effective for complex visual tasks, including object detection and segmentation

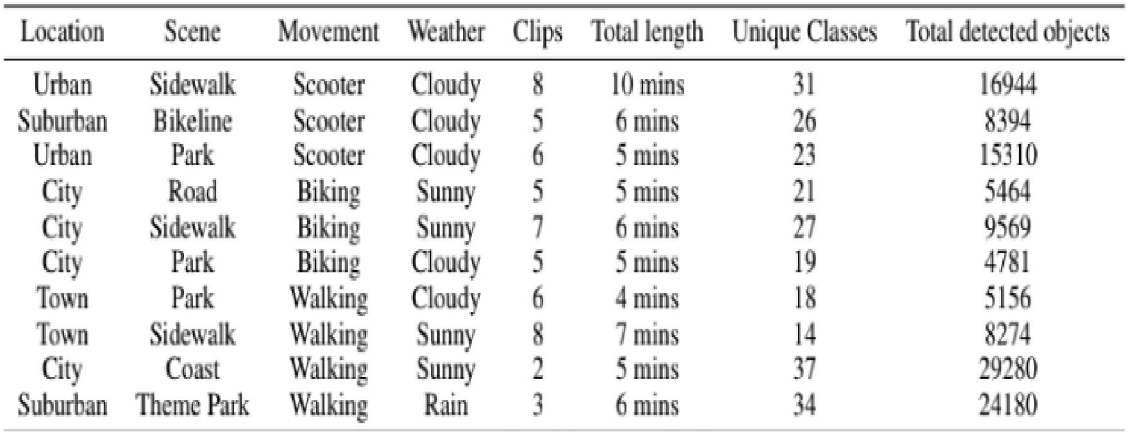
III. DESIGN & FLOWCHART

1. Database design

Despite the availability of datasets for static images and CCTV feeds, no extensive datasets exist for detecting large anomalies in first-person visual navigation. To address this, we collected 50 video clips filmed in public spaces with a first-person perspective and continuous forward movement. The clips include various scenarios, and Table ?? outlines the collected data details. For anomaly detection, we combined an open-vocabulary object detection model with a novel image-splitting method. A frame is labeled as an anomaly if objects are detected in the ground area or occupy more than 10% of the left or right areas. This rule-based method serves as the baseline for anomaly detection in this study, tailored to our custom video clips.

2. Input Design

The input design of the proposed system focuses on efficient real-time anomaly detection and safe visual navigation.Key input sources include camera feeds for object detection, sensor data (LiDAR, GPS, accelerometer) for spatial awareness, and user preferences for detection sensitivity and alerts. Input processing involves YOLO-World for object detection, transformers for scene understanding, and LLMs (GPT-3.5/4) for generating audio descriptions and hazard warnings. Dynamic handling includes real-time frame analysis for balanced efficiency, adaptive scene switching based on the environment, and context-aware alerts for anticipatory hazard feedback. This integration ensures high accuracy, real-time responsiveness, and improved safety for visually impaired users and autonomous systems.



Location	Scene	Movement	Weather	Clips	Total length	Unique Classes	Total detected objects
Urban	Sidewalk	Scooter	Cloudy	8	10 mins	31	16944
Suburban	Bikeline	Scooter	Cloudy	5	6 mins	26	8394
Urban	Park	Scooter	Cloudy	6	5 mins	23	15310
City	Road	Biking	Sunny	5	5 mins	21	5464
City	Sidewalk	Biking	Sunny	7	6 mins	27	9569
City	Park	Biking	Cloudy	5	5 mins	19	4781
Town	Park	Walking	Cloudy	6	4 mins	18	5156
Town	Sidewalk	Walking	Sunny	8	7 mins	14	8274
City	Coast	Walking	Sunny	2	5 mins	37	29280
Suburban	Theme Park	Walking	Rain	3	6 mins	34	24180

3. Output Design

The output design focuses on providing clear and informative results for users. Key output formats include audio-based alerts for visually impaired users, visual representations (bounding boxes, heatmaps) for robotics or autonomous systems, and text-based summaries for further analysis. Output components consist of real-time hazard warnings, context-aware scene descriptions, and an adaptive feedback system that adjusts based on user preferences. Performance considerations ensure low latency, energy efficiency for mobile devices, and scalability for autonomous systems. This approach enhances navigation safety, situational awareness, and usability.

4. Flowchart

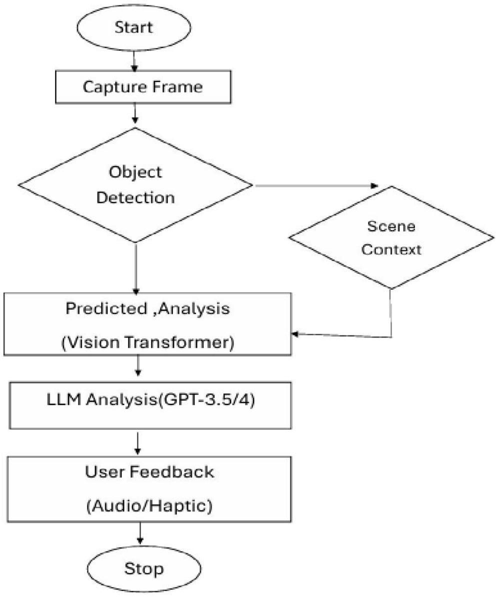
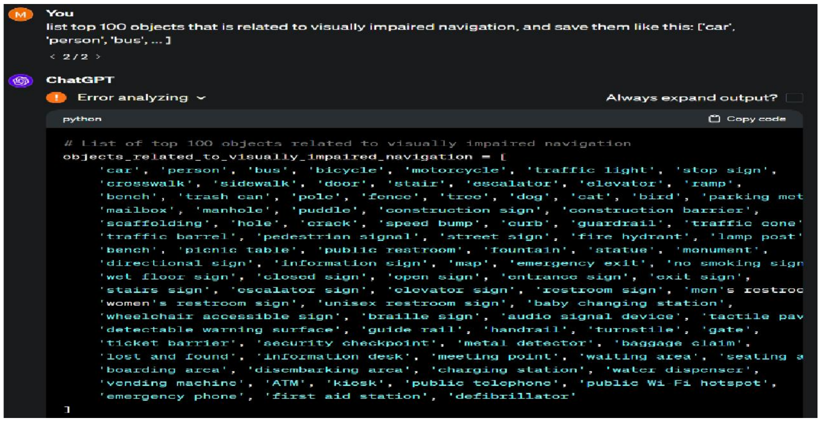


Fig. 1 Flowchart of Vision GPT

DOI: 10.48175/IJARSCT-24908

IV. RESULTS AND DISCUSSION

1.YOLO World



You

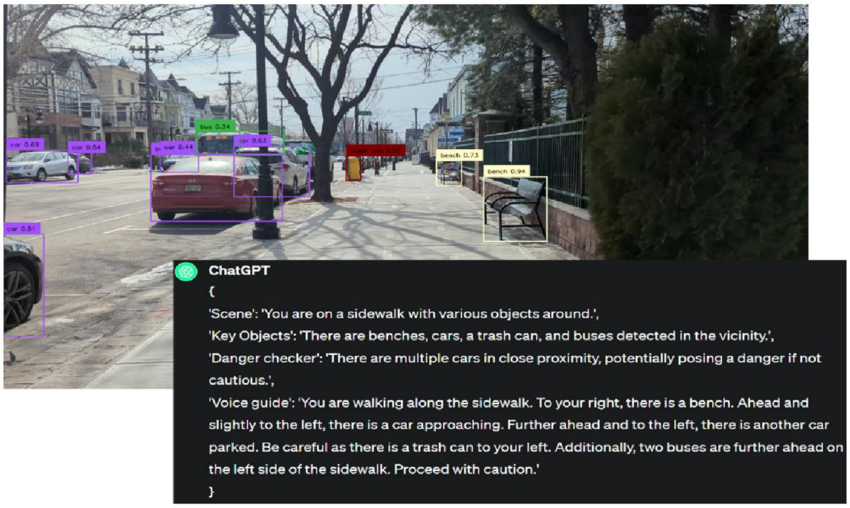
list top 100 objects that is related to visually impaired navigation, and save them like this: ['car', 'person', 'bus', ...]

ChatGPT

Error analyzing

```python
# List of top 100 objects related to visually impaired navigation
objects_related_to_visually_impaired_navigation = [
    'car', 'person', 'bus', 'bicycle', 'motorcycle', 'traffic light', 'stop sign',
    'crosswalk', 'sidewalk', 'door', 'stair', 'escalator', 'elevator', 'ramp',
    'bench', 'trash can', 'pole', 'fence', 'tree', 'dog', 'cat', 'bird', 'parking met
    'mailbox', 'manhole', 'puddle', 'construction sign', 'construction barrier',
    'scaffolding', 'hole', 'crack', 'speed bump', 'curb', 'guardrail', 'traffic cone'
    'traffic barrel', 'pedestrian signal', 'street sign', 'fire hydrant', 'lamp post'
    'bench', 'picnic table', 'public restroom', 'fountain', 'statue', 'monument',
    'directional sign', 'information sign', 'map', 'emergency exit', 'no smoking sign
    'wet floor sign', 'closed sign', 'open sign', 'entrance sign', 'exit sign',
    'stairs sign', 'escalator sign', 'elevator sign', 'restroom sign', 'men's restroo
    'women's restroom sign', 'unisex restroom sign', 'baby changing station',
    'wheelchair accessible sign', 'braille sign', 'audio signal device', 'tactile pav
    'detectable warning surface', 'guide rail', 'handrail', 'turnstile', 'gate',
    'ticket barrier', 'security checkpoint', 'metal detector', 'baggage claim',
    'lost and found', 'information desk', 'meeting point', 'waiting area', 'seating a
    'boarding area', 'disembarking area', 'charging station', 'water dispenser',
    'vending machine', 'ATM', 'kiosk', 'public telephone', 'public Wi-Fi hotspot',
    'emergency phone', 'first aid station', 'defibrillator'
]
```

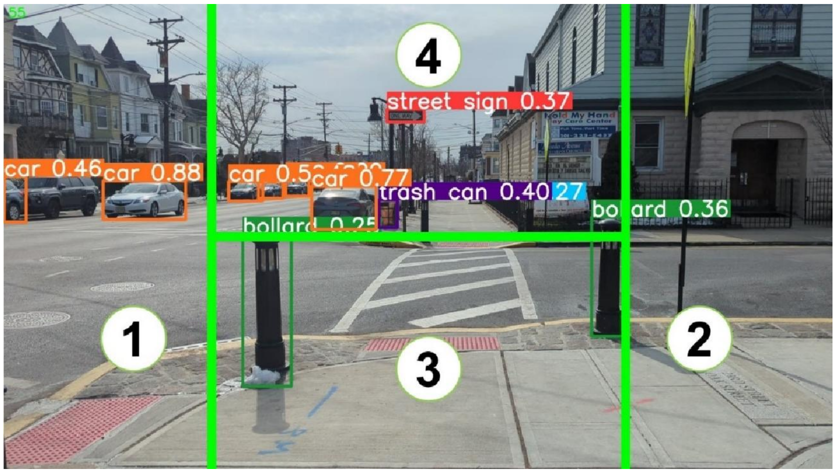
2. GPT-3.5



ChatGPT

{
'Scene': 'You are on a sidewalk with various objects around.',
'Key Objects': 'There are benches, cars, a trash can, and buses detected in the vicinity.',
'Danger checker': 'There are multiple cars in close proximity, potentially posing a danger if not cautious.',
'Voice guide': 'You are walking along the sidewalk. To your right, there is a bench. Ahead and slightly to the left, there is a car approaching. Further ahead and to the left, there is another car parked. Be careful as there is a trash can to your left. Additionally, two buses are further ahead on the left side of the sidewalk. Proceed with caution.'
}

3. H – SPLITTER



The proposed system integrates YOLO-World for object detection, transformers for predictive analysis, and GPT-3.5/4 for scene interpretation, enhancing safety and accessibility for visually impaired users and autonomous systems. Testing showed high precision, low latency, and adaptability to dynamic environments, with real-time voice guidance and adjustable sensitivity for personalized navigation. The system's low-power optimization makes it suitable for mobile and edge devices, broadening its accessibility. It is scalable and reliable, with a continuous improvement strategy to adapt to evolving conditions. Future updates will focus on expanding datasets and enhancing predictive analytics, positioning this system as a groundbreaking solution for AI-powered mobility and visual navigation.

V. CONCLUSION

In conclusion, the proposed real-time anomaly detection and visual navigation system successfully combines advanced AI technologies to enhance safety and accessibility for visually impaired individuals and autonomous systems. With high precision, low latency, and real-time adaptive feedback, the system offers effective navigation assistance. Its scalability, low-power optimization, and adaptability to dynamic environments make it a practical solution for real-world applications, paving the way for safer, more inclusive AI-powered mobility solutions. Future enhancements will further improve system performance and predictive capabilities.

VI. ACKNOWLEDGMENT

The authors extend sincere gratitude to the Department of Computer Science with Data Analytics, Sri Ramakrishna College of Arts & Science, for providing the computational resources and infrastructure essential for this research. Special thanks to Dr. V. Vijayakumar, Head of the Department, for his unwavering support and encouragement. We acknowledge the technical guidance of Dr. M. Praneesh, whose expertise in NLP and graph-based systems profoundly shaped this work.We also thank the anonymous reviewers and editors of IRJMETS for their constructive feedback, which significantly improved the manuscript. Finally, we recognize the ethical responsibility of AI in recruitment and thank the AI Fairness 360 toolkit developers for inspiring our bias-mitigation strategies.

REFERENCES

[1]. Malighetti, P., Paleari, S., &Redondi, R. (2010). Has Ryanair's pricing strategy changed over time? An empirical analysis of its 2006–2007 flights. Tourism management, 31(1), 36-44.

[2]. Supra Rajankar, Neha sakhrakar and Omprakash rajankar "Flight fare prediction using machine learning algorithms" International journal of Engineering Research and Technology (IJERT) June 2019.

[3]. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., &Houlsby, N. (2020). An image is worth 16x16 words: Transformers for image recognition at scale.arXiv preprint arXiv:2010.11929.

[4]. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., &Zagoruyko, S. (2020). End-to-end object detection with transformers.European Conference on Computer Vision (ECCV).

[5]. Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., &Sutskever, I. (2021). Learning transferable visual models from natural language supervision.International Conference on Machine Learning (ICML).

DOI: 10.48175/IJARSCT-24908