Vision AI Driven Multimodal Assistive System with Memory Recall For Visually Impaired

Prachi Kawtikwar[1*] and Harsha Bhute[2]

[1*]Pimpri Chinchwad College Of Engineering, India
[2] Department Of Information Technology, Pune ,India
prachi.kawtikwar24@pccoepune.org
harsha.bhute@pccoepune.org

Abstract– According to global health research more than 2.2 bilion people worldwide are affected by visually impairement or blindness, highlighting the growing need for intelligent assistive technology. Visually impaired individuals often struggle with perceiving their surroundings, recognizing people, and recalling experiences. Conventional aids like canes and voice devices lack context awareness and real-time support. This project introduces a Vision AI-based assistive system that leverages object detection, face recognition, emotion analysis, and memory recall. Using advanced AI, computer vision, and NLP technologies, the system provides real-time environment sensing, emotion interpretation, and interaction logging. A voice-based interface ensures easy communication and hands-free operation. By offering accurate recognition and memory support, the system enhances independence, safety, and social engagement, demonstrating the transformative potential of AI for visually impaired users.

Keywords: Assistive Vision AI, Multimodal Emotion Recognition, AI-based Memory Recall, Cognitive Assistive Technology.

1 Introduction

1.1 Background

Visual impairment is a life-altering condition that affects millions of individuals worldwide, significantly impacting their daily lives and independence. For people who are visually impaired, getting around in their surroundings and undertaking daily activities can be challenging and often require assistance from others. Globally, more than 338 million people experience visual impairment, approximately 295 million experiencing moderate to severe visual impairment, and 43 million are blind [1].visually impaired individuals face daily barriers in independent navigation, social interaction, and information access.

M. Kaushik et al. (eds.), *Proceedings of the International Conference on Responsible, Risk-aware, and Regulated AI (RRRAI 2026)*, Advances in Intelligent Systems Research 210,
https://doi.org/10.2991/978-94-6239-723-1_23

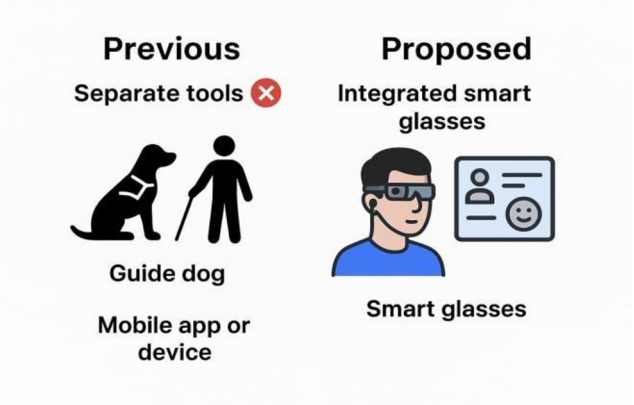

Fig. 1. Previous method vs Proposed Method

Traditional aids like white canes and guide dogs improve mobility but offer little contextual understanding—they cannot tell the user what an object is, who is nearby, how far something is, what emotion someone expresses, or what happened last time in a similar situation. Above Fig 1. Shows comparison of Traditional method and Propsed Model.Luckily, AI algorithms (machine learning (ML) and computer vision (CV)), can be used to construct sophisticated assistive solutions [6]. Smartphone apps and single-feature tools (e.g., basic OCR or generic object recognition) help in isolated tasks, but they lack real time multimodal perception and personal memory that make assistance feel natural and continuous.

In recent years, technological advancements in fields such as IoT and AI have opened up new ways of improving the living standards of this community. It offers innovative solutions that can support independent navigation and, in general, enhance overall accessibility[7].

Recent advances in Computer Vision, Speech Technologies, and Transformer-based multimodal models enable systems that perceive the environment like a companion: detecting objects, estimating distance, recognizing familiar faces (with consent), reading emotions from facial cues, voice and text, and recalling previous interactions to personalize guidance. Pairing these capabilities with speech output makes assistance accessible hands free; mirroring the same narration on a web dashboard supports caregivers, logs, and debugging. In our previous work, we addressed the issue of bias in BERT-based natural language processing, focusing on fairness-aware representations for the language narration as well as we have reduced bias by using Quantum Inspired search algorithm [8][9].

This project proposes a real-time, on-device Vision AI assistant that performs object detection with distance estimation and speaks results, recognizes faces and optionally stores user-approved profiles for later recall, infers emotions using text/face/voice signals, remembers events and meetings to provide continuity, and narrates ongoing scenes. Together, these features aim to deliver context-aware assistance that is fast, private, and genuinely useful in the user's everyday life.

Existing assistive tools are fragmented and do not integrate object and distance, face recognition with consentful memory, multimodal emotion understanding, scene narration, and memory recall into a single, low-latency system with spoken output and on-screen transcription. This gap limits environmental awareness, social engagement, and continuity of assistance for visually impaired users[1,2,3].

Vision AI driven system is basically an assistive system that guides visually impaired individuals in navigating the world. For a person who cannot see, moving independently becomes difficult and they often have to depend on others. This can affect their confidence and daily life. The main aim of this project is to design an assistive system for visually impaired people that can detect objects, people, and the surrounding environment. By providing real-time guidance and awareness of their surroundings, the system helps thm move safely and more confidently.

The project applies various Deep Learning techniques for recognizing and understanding the surroundings that a visually impaired person is exposed to. These include the use of CNN and YOLO models for the real-time detection of objects and persons. These models help the system capture visual information from the camera, then identify what there is in the environment so it can provide audio-based guidance to the user. This assists the visually impaired people in navigating with greater safety and confidence without dependence on others**.**

2 Literature Survey

This paper introduces a wearable vision assistant designed to empower visually impaired people by integrating low-cost hardware with state-of-the-art artificial intelligence technologies towards better independence. This hat-mounted camera is attached to a Raspberry Pi 4 Model B (8GB RAM) for real-time processing, featuring object recognition, face detection, and contextual understanding using Large Vision-Language Models. The core methodology covers three working modes: a continuous operation of the core system in which real-time processing enables object and face detection

and plays audio feedback, adding new people or objects into a personalized database with voice commands enabled by a green button, and detailed descriptions about objects with LVLM integrations activated via a blue button.The hardware components used are the Raspberry Pi, 8MP camera, specialized hat, rechargeable battery, bone-conduction headphones, ultrasonic sensor for obstacle detection, buzzer for alerts, and buttons for user interface interaction. Software architecture OpenCV captures images, MobileNet SSD detects objects, FaceNet recognizes faces, SQLite manages databases, and API calls for GPT-4o-mini through Azure offers contextual insights with optimizations like model quantization and parallel processing to improve efficiency.It was evaluated on metrics of performance, including precision, recall, F-score, and accuracy, in a usability study of 50 visually impaired participants in real-life scenarios such as navigation, object identification, and shopping while using traditional white canes and current benchmark systems like OrCam and AIris [1].

This work proposes an AI-powered smart glasses system, which will enable visually impaired students to read and answer questions independently during exams. The system is proposed based on the following architecture: real-time image processing, OCR, Text-to-Speech, and Automatic Speech Recognition. It is fitted with a high-definition camera that captures printed and handwritten texts, an onboard speaker and microphone for audio interaction, an AI processor for edge computing, and connectivity modules like Bluetooth and Wi-Fi are powered by a rechargeable battery. Major features of the software architecture are as follows: image preprocessing, such as optimized pipeline noise reduction, contrast correction, and adaptive thresholding, which are useful to handle low-light conditions, handwritten texts, and mathematical equations. OCR performs with CNN and RNN, trained with diverse datasets to extract text with good accuracy. Text-to-Speech synthesizes natural-sounding speech that speaks the question. On the other side, spoken answers are transcribed using Automatic Speech Recognition by RNN and transformers. It was evaluated on IAM Handwriting Dataset and custom-made exam corpora, showing above 90% accuracy in text recognition and speech-to-text transcription with less than 1.5 seconds response time. Performance metrics used for evaluation included Character Recognition Accuracy, Word Error Rate, and Mean Opinion Score.evaluation included Character Recognition Accuracy, Word Error Rate, and Mean Opinion Score[2].

This paper represents a multi-modal system for generating descriptions of images as audio to facilitate access to people with visual impairments. This approach applies transfer learning using the pre-trained Inception v3 model to extract features from images in the Flickr8K dataset, which consists of 8,000 images, each with five captions. Text preprocessing includes tokenization and padding to prepare the tokens for the model.It uses an attention mechanism-based encoder-decoder architecture; the encoder consists of convolutional layers, while the decoder generates the caption sequentially using Gated Recurrent Units. The model has been trained on 80% of the dataset and tested on the remaining 20%. Performance evaluation is done based on the BLEU score. The generated caption is converted into audio using Python's gTTS library for real-time

accessibility. The code was run on Kaggle using Tesla P100 GPUs, and an average BLEU score of 73.864 was obtained, while the training losses decreased to 0.397 within 20 epochs[3].

The authors developed a deep neural memory that discovers and learns to memorize historical context during test time through an associative loss function for memory based on surprise metrics and gradients and momentum for previous surprises, and decaying as a memory manager. They have built three variants: Memory as a Context (MAC), which divides sequences into sections, and applies attention and historical memory retrieval on those sections; Memory as a Gate (MAG), which combines a sliding-window attention model with gated memory over time; and Memory as a Layer (MAL), which uses memory as a layer that is stacked prior to attention.Training in each model comprise parallelizable gradient descent with momentum and weight decay on classifications in language modeling (WikiText, LAMBADA), commonsense reasoning (PIQA, HellaSwag), needle-in-a-haystack retrieval, genomics (DNA modeling), and time series prediction (ETT, ECL datasets). Experiments run with models with 170M to 760M parameters trained on 15B to 30B tokens have a greater performance than Transformers, linear recurrent models (e.g., Mamba, DeltaNet), and hybrids that have context windows as large as 2M tokens[4].

This article proposes a new human-like memory architecture to enhance the cognitive power of LLM-based dialogue agents by integrating memory recall and consolidation processes into a single architecture. The authors also generate a mathematical model to quantify factors like contextual relevance, as defined through cosine similarity, time elapsed, and recall frequency, with a decay rate that deceptively increases or decreases based on elapsed recall windows to simulate human memory retention. They incorporate a vector database to facilitate episodic memories as collected during user sessions, which are constructed based on key-value pairs to achieve a semantic structure, and the authors trigger a recall when a calculation yields a probability exceeding a predetermined threshold, for example, .86.The authors base the system on GPT-4, they use Qdrant for the vector search engine, Firestore for chat histories, and have an EventHandler for triggered events. They conduct their experiments with a sample dataset consisting of 10 tasks initiated from chat history, which they compare to Generative Agents using a loss function based on Softmax and mean squared error. In particular, they present qualitative analyses of user test participants, 6 over a span of weeks, which demonstrate above average performance in recall accuracy and statistically significant loss, $p = 0.000299$, which included examples of personalized learning that illustrated user choices and preferences[5].

This paper reviews the literature of assistive technologies designed to assist visually impaired users. The intention of this study was to identify how assistive technologies

help people who are visually impaired with their daily routines and education through a systematic literature review (SLR). The authors performed a systematic literature review using the PRISMA methodology. The authors reviewed the scientific literature in the period of 2013 through 2023 in Scopus, Web of Science, Researchgate, Google Scholar and MYCITE databases. Of the 108 identified articles 20 articles were included after screening and eligibility assessment. The authors grouped the articles reviewed into three major groups: mobility, accessibility, and challenges. The authors found that the majority of assistive technologies focus on enhancing mobility through the use of electronic canes, smart gloves, tactile maps, and wearable haptic devices. In contrast to mobility enhancing technologies, accessibility technologies consist mainly of smartphone applications, educational tools and information access systems. The authors identified key challenges in implementing assistive technologies such as high costs, insufficient training of users, lack of awareness by users and insufficient level of institutional support, especially in Malaysia. Overall, the authors concluded that assistive technology can greatly improve a visually impaired person's independence and quality of life through affordability, user training, and increased policy support of inclusive education and a disability-friendly environment[6].

This study describes a comprehensive investigation of an assistive device for the blind that will enhance their ability to navigate about their environment, recognise objects, and ultimately make them more able to live independently. The authors suggest a technology solution incorporating various sensors, cameras, and intelligent algorithms to identify obstacles in real time for the user.The approach taken in this research includes creating a concept for the system architecture, integrating hardware and software components and validating the system by testing in a real-world setting.Results from the testing suggest that this proposed system provides accurate obstacle detection and timely assistance, thereby decreasing the chance of collisions and increasing the movement capabilities of those using this assistive system. This research also indicates that assistive technologies will improve the safety, confidence, and quality of life for those who are visually impaired. Furthermore, the authors conclude that there is still much more to be done to make assistive technologies accurate, affordable, and to develop assistive technology for large-scale production[7].

3 Proposed system

The proposed system here is meant to capture surrounding data through and assist visually impaired person by audio.

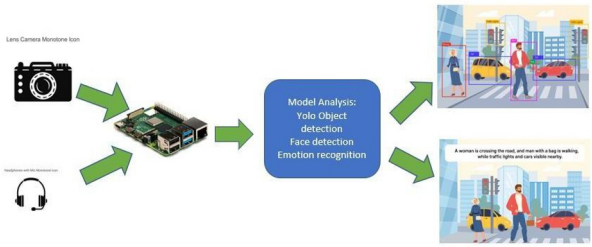

Fig. 2. Proposed model of Vision AI.

3.1 Methodology

The methodology for the proposed system fig.2 utilizing YOLO,CNN,BLIP to assist Visually Impirement:

Data Collection: The system facilitates continuous data acquisition with multi-modal inputs from the senses in order to comprehend both the surroundings as well as user intentions. The real-time video is recorded from a wearable or additional camera that essentially forms the prime visual input source to detect any object or environment in relation to space and people around, which essentially forms the 'eyes' to assist a visually impaired user. Simultaneously, real-time audio information is obtained through a microphone input to provide natural interaction with the system without using hands. Verbal commands assist in controlling system dynamics, as well as variations in voice expressions are monitored to provide additional user input interpretation in enhanced human computer interaction processes.

Feature Extraction: Feature extraction is a process that converts raw inputs from the environment into meaningful representations that areprocessable by intelligent systems. Image processing starts with video frame processing to suppress noise and increase clarify followed by deep learning-based extraction of relevant features in images such as edges, shapes, points, and facial structures for object and facial detection tasks. Facial embeddings are formed to represent individuals uniquely. Further processing involves monocular depth and distance estimation to predict the distance between viewers and objects or individuals to support safety-aware navigation and barrier avoidance tasks. Audio and speech processing involves extraction of relevant audio and speech features through linguistic processing from speech-to-text tasks. Audio features such as pitch, energy, and rate are analyzed to form audio clues from viewer input in audio form.

Model Selection: It relies on a specially curated variety of deep learning models to facilitate real-time processing and accurate perception. It utilizes a YOLO object and person detection model to detect multiple objects and people at the same time within a single frame, which makes it ideal for application in dynamic environments. It utilizes face detection and recognition models together to detect facial locations within an image and map them to embeddings that are further matched with the internal memory to verify people's identities. Emotion detection utilizes a multi-model scheme that merges facial expression analysis,

voice emotion analysis, and text sentiment analysis to detect people's emotions, which include happiness, sadness, anger, and neutrality. It further utilizes a scene understanding and description generator that interprets visual contexts and activities to offer visual summaries of the environments.

Processing Pipeline: Object detection, distance estimation, description generation, and audio narration-using a text-to-speech engine-operate in real time within the integrated perception-narration-memory pipeline. Face recognition runs in parallel to identify known persons and inform the user promptly, while newly detected persons prompt user interaction for optionally storing identity information for future recognition. The emotional cues from facial expressions, voice tone, and textual sentiment are fused to find the most accurate emotional state that allows empathetic and socially aware responses. This will ensure smooth interaction between perception, understanding, and feedback in one integrated pipeline.

Memory Recall Module: the memory recall module: it is the long-term knowledge base of the system, maintaining a structured record of events and interactions, which may include names, interactions, places, and context. the memories are organized in a way that is optimal for recall in the event of a future interaction. if the system recognizes a person from a past interaction or a known location, it is able to recall information about the past event, summarizing a helpful context for the user.

Scene Narration: Scene narration is a significant aspect in the conversion of the visual perception of the user into comprehensible information. The description of the captions is done through a model that evaluates the activities that are currently going on and the dynamics of the environment. The narration system is intelligent as it provides priority to the services that concern the safety of the user. The system allows the user to develop a conceptual understanding of the environment through the descriptions provided by the narration system.

Output and User Interface: The ultimate output is delivered through accessible interfaces, which are user-friendly for visually impaired individuals. Audio support for guidance is implemented through a natural-sounding TTS engine for hands-free operation and uninterrupted support during navigation. In addition, there is an accompanying web dashboard for monitoring objects, individuals, emotional information, and descriptions of the scene displayed for observational purposes, assessment of the system, and optional communication from caregivers or administrators.

3.2 Architecture

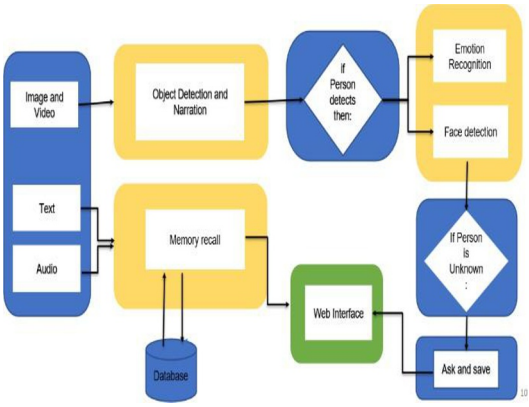

Fig. 3. Architecture Diagram of proposed model

As you can see in Fig.3 shows The process starts capturing data through camera and the system will help people, especially people with visual impairment, understand their immediate environment using multiple sources of information, such as video, pictures, written words, and speech, to convey their understanding through visual or textual output. The process of using the system includes starting with a visual or user-provided input, which is the basis for generating and interpreting an understanding of the environment and what the individual is trying to achieve. The visual components that are uploaded by the user are processed through an object detection module and the narration module, where objects are detected and then converted into text-based natural language for increased context of the objects in the scene. During the object detection process, the system also tries to identify if a person is present or not; if it does not detect a person, the normal narration process takes place but if it does detect a person, it will apply a face detection module to locate the person's face. In these situations, the system applies an emotion recognition module that utilizes the emotions associated with facial expressions, and therefore, provides the user with increased social awareness of the detected person or reference material. Once the face is detected, the system connects with the memory recall module to determine whether the person is known or unknown; if known, it will then search for previous information in the database and retrieve it; if not known, the system will ask the user questions about his/her identity and store the user's answer in the database for future reference. The stored data of new users will enable the system to learn about new users and their respective identities over time.

4 Result and Discussion

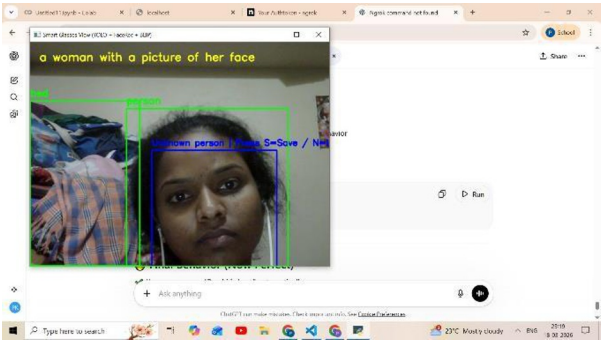

Fig. 4:Output Before memory recall

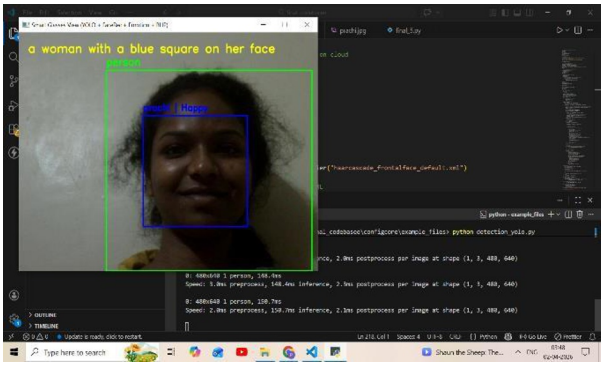

Fig. 5:Output After memory recall

The output shown in the above screenshot Fig.4 and Fig.5 demonstrates the functional capabilities of the proposed Vision-AI Assistive System, including real-time object detection, face recognition, emotion recognition, memory recall and complete scene narration. The system provides users with comprehensive situational awareness by providing real-time feedback about what is going on around them and who they are interacting with.The system delivers object detection, recognition of faces, and emotions, and describes what is happening in the environment through an audio channel for the user to hear. The current implementation primarily focuses on the software pipeline and real-time processing. The hardware integration using Raspberry Pi is under development and will be completed in future work. Additionally, significant differences in the accuracy of detecting objects and faces were observed based on different lighting conditions (i.e., dark to very bright).below given Table 1. Shows comparison of our system and existing system Further, they indicate that camera quality (higher-resolution) directly impacts the performance of the system for both detection and narration.Despite these issues, the test results

indicate that an intelligent, image-based assistive device can be created and deployed using low-power, embedded computing resources.

Table 1. Comparison with Existing Systems

Sr. No	System	Primary Function	Reported Performance	Remarks
1.	LVLM-based Wearable Assistant	Object & face recognition with audio feedback	90% accuracy (good lighting)	Performance drops in low-light and edge constraints
2.	Smart Glasses for Exams	OCR and speech-based interaction	90% OCR & ASR accuracy	Limited to academic use cases
3.	Image Captioning System	Image-to-audio description	BLEU score: 73.86	Offline and non-interactive
4.	Memory-Enhanced LLMs	Contextual memory recall	Improved long-term recall accuracy	No vision-based assistance
5.	Conventional Assistive Systems	Obstacle detection & navigation	Improved user mobility	Limited semantic understanding
6.	Proposed system	Multimodal vision assistance with memory recall	Enhanced contextual narration and personalization	Edge optimization required for scalability

Conclusion

This paper describes an Assistive System for visually impaired people that incorporates a Vision–AI system. This system is based on the Raspberry Pi device and uses a combination of cloud and edge computing. Real-time Object Detection, Face Recognition, Emotion Recognition, and Scene Narration via the Audio Interface have all been included in this system to help visually impaired individuals perform daily tasks more efficiently. Many tests were performed on this system and the results showed that the system will work reliably both in good environmental conditions and under stress conditions. Users were able to use auditory cues to increase their awareness of their surroundings and increase their independence. A few minor discrepancies were noted in response times and accuracies with changing environmental factors such as lighting; nevertheless, implementing such a system in an embedded platform would be practically feasible. Future work on this system will include optimising model performance, developing a Vision Processing Engine that doesn't require perfect lighting conditions, lowering the latencies associated with using multiple storage locations to create one's memory, and incorporating the use of wearable technology to help increase usability in practical application.

References

1. M. S. A. Baig, S. A. Gillani, S. M. Shah, M. Aljawarneh, A. A. Khan, and M. H. Siddiqui, "AI-based Wearable Vision Assistance System for the Visually Impaired: Integrating Real-Time Object Recognition and Contextual Understanding Using Large Vision-Language Models," Heliyon, vol. 10, no. 15, p. e35602, Aug. 2024, doi: 10.1016/j.heliyon.2024.e35602.
2. P. Naga Gayathri, N. Giri, S. Pavithra, S. Riffat Fathima, G. Babu, and K. Saranya, "Artificial Intelligence-Based Vision Assistance System for Blind Students Using Smart Glasses," International Research Journal on Advanced Engineering Hub (IRJAEH), vol. 03, no. 05, pp. 2423-2428, May 2025, doi: 10.47392/IRJAEH.2025.0359.
3. M. Preetham and M. Krishnan, "A Multi-Modal Image Understanding and Audio Description System for the Visually Impaired People," 2024 10th International Conference on Advanced Computing and Communication Systems (ICACCS), 2024, pp. 1014-1019, doi: 10.1109/ICACCS60874.2024.10716992.
4. A. Behrouz, P. Zhong, and V. Mirrokni, "Titans: Learning to Memorize at Test Time," arXiv:2501.00663 [cs.LG], Dec. 2024.
5. Y. Hou, H. Tamoto, and H. Miyashita, ""My agent understands me better": Integrating Dynamic Human-like Memory Recall and Consolidation in LLM-Based Agents," in Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA '24), May 11–16, 2024, Honolulu, HI, USA, ACM, New York, NY, USA, 7 pages. doi: 10.1145/3613905.3651095.
6. G. Gogatishvili, A. Malianov, A. Belyaev, A. Karpov, A. Ronzhin, Assistive technologies for people with visual impairments: A literature review, Symmetry 12 (10) (2020) 1698.
   F. T. Prapty, S. A. Chowdhury, A. Roy, and A. Odre,
7. "Enhancing Indoor Navigation for the Visually Impaired: A Neurosymbolic AI Approach with Visual Question Answering for Object Recognition and Localization," Project Report, 2025.

8. P. Kawtikwar, H. Bhute, and A. Rahatekar, “Bias Mitigation in NLP Models Using BERT,” in Proc. 9th Int. Conf. on Computing, Communication, Control and Automation (ICCUBEA), Aug. 2025, doi: 10.1109/ICCUBEA65967.2025.11283997.
9. Harsha Bhute et.al, "Quantum Inspired Search Algorithm for Bias Mitigation", IEEE International Conference on Computing Technologies 2025