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 6, Issue 9, March 2026
Impact Factor: 8.2

VisionTalk: AI-Based Visual Understanding and Speech Interaction System

Siva Kumar Mondru[1], S Mani Kuchibhatla[2], Tanisha Tompala[3], Vinuthna Pagolu[4], Hemanth Velpula[5]
Assistant Professor, Department of Information Technology, ACE Engineering College, Hyderabad, India [1]
Professor & HOD-IT, Department of Information Technology, ACE Engineering College, Hyderabad, India [2]
Student, Department of Information Technology, ACE Engineering College, Hyderabad, India [345]

Abstract: *Vision-Talk is an AI-based system that helps users identify and understand objects through camera input and voice interaction. The system allows a user to point the camera at an object and ask questions such as "What is this?" or "How do I use it?". The object is detected in real time using the YOLOv8 model implemented with Python. The application interface is built with React and Vite to deliver a fast, responsive user experience. After detecting the object, the system generates a simple explanation and converts the response into speech using gTTS (Google Text-to-Speech). This allows the user to hear the explanation clearly. By combining computer vision, speech technology, and a modern web interface, Vision-Talk provides an easy and interactive way for users to understand objects around them*

Keywords: Artificial Intelligence, YOLOv8, React, Vite, Python, gTTS, Object Detection

I. INTRODUCTION

Artificial Intelligence (AI) has transformed human-computer interaction by enabling machines to understand visual information and respond to user queries. Recent advancements in computer vision and multimodal AI models have enabled systems to recognize objects, interpret scenes, and generate contextual explanations.

Modern smartphones and digital assistants such as Apple Intelligence and Google Gemini provide visual recognition features that allow users to identify objects using camera input. While these systems can provide object descriptions, they typically function as general-purpose assistants and do not focus on procedural guidance or task-specific assistance.

In real-world scenarios, many users require more than object recognition. For example, individuals may need assistance in understanding how to operate unfamiliar devices, perform repairs, or follow safety procedures. Existing systems primarily provide information retrieval rather than structured guidance.

To address this limitation, this paper proposes VisionTalk, an AI-based system that combines visual recognition, contextual reasoning, and speech interaction to assist users in understanding objects and performing tasks. The system integrates camera input with multimodal AI models capable of interpreting visual information and generating human-readable explanations.

Copyright to IJARSCT
www.ijarsct.co.in
DOI: 10.48175/IJARSCT-32713

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 6, Issue 9, March 2026
Impact Factor: 8.2

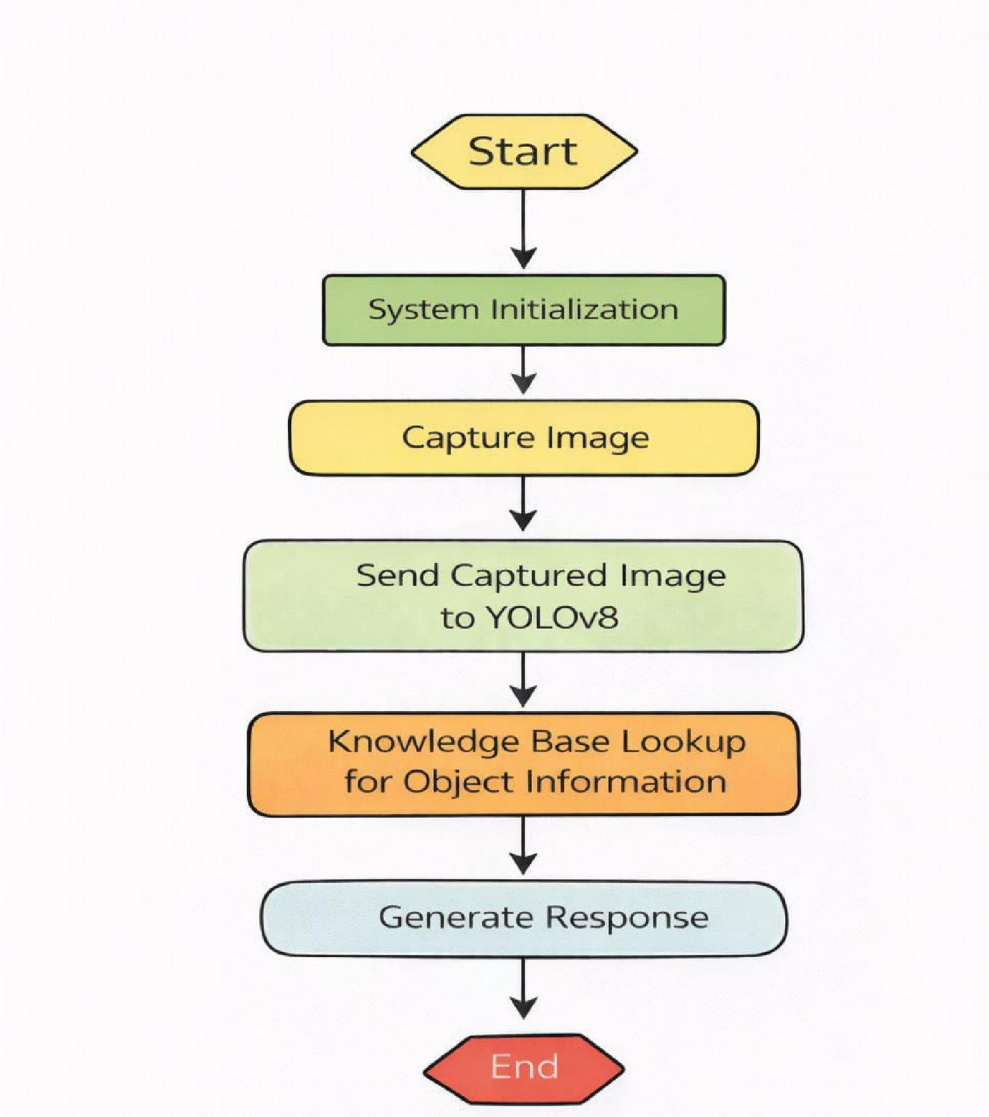

Fig. 1 Image Interaction Interface of VisionTalk

The primary objective of VisionTalk is to develop an interactive AI guidance system capable of identifying objects and providing step-by-step assistance using natural language and speech interaction. This system can be applied in domains such as education, maintenance assistance, and accessibility support.

II. LITERATURE REVIEW

Recent advancements in artificial intelligence have enabled the development of intelligent systems capable of understanding visual environments and generating contextual explanations. Computer vision and multimodal learning have played a significant role in enabling machines to interpret images and interact with users through natural language.

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 6, Issue 9, March 2026
Impact Factor: 8.2

DetGPT introduced a reasoning-based object detection approach that integrates vision models with language models to detect objects using natural language instructions. Instead of relying solely on predefined object categories, DetGPT allows users to query objects through language, enabling flexible and intelligent visual reasoning [1].

Visual Question Answering (VQA) systems represent another important advancement in multimodal AI. These systems combine image understanding with natural language processing to answer questions related to visual scenes. Antol et al. proposed a VQA framework that allows machines to analyse images and respond to user queries using contextual reasoning [2].

Object detection models such as YOLO have also played a significant role in real-time computer vision systems. YOLO-based models are widely used due to their ability to detect multiple objects in real-time with high speed and accuracy, making them suitable for interactive applications [3].

Several assistive technologies have also been developed for visually impaired individuals. These systems typically detect surrounding objects using cameras and provide audio descriptions to help users understand their environment. However, most of these systems focus only on object identification rather than interactive guidance or contextual explanations [4].

Although existing research demonstrates the capability of object detection and multimodal reasoning, most systems are limited to providing descriptive information. They do not offer structured procedural guidance or interactive assistance. The proposed VisionTalk system addresses this limitation by integrating object detection, multimodal reasoning, and speech interaction to provide contextual guidance to users.

III. METHODOLOGY

The proposed VisionTalk system is designed as an intelligent visual understanding and speech interaction platform that assists users in identifying objects and understanding how to use them through real-time AI guidance.The overall working of the VisionTalk system is illustrated in Fig. 2

The system integrates computer vision, multimodal reasoning, and speech interaction to create an interactive environment between the user and the AI system. The overall architecture consists of five main modules:

- Camera Input Module
- Object Detection Module
- Multimodal AI Reasoning Module
- Speech Interaction Module
- User Interface Module

The camera input module captures real-time video frames from the user's device. These frames are processed and converted into image data suitable for AI-based analysis.

The object detection module utilizes pretrained object detection models such as YOLO to identify objects present in the captured image. The model detects objects and provides labels along with bounding boxes.

Once the object is detected, the system forwards the visual information to the multimodal reasoning module. This module uses advanced AI models such as Gemini to analyze the object and generate contextual explanations. The reasoning module can interpret the object's purpose, usage, and safety considerations.

The speech interaction module converts the generated text response into speech output. This enables users to interact with the system through voice-based responses, improving accessibility and usability.

Finally, the user interface module displays detection results, interaction logs, and contextual guidance to the user through a web-based interface.

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 6, Issue 9, March 2026
Impact Factor: 8.2

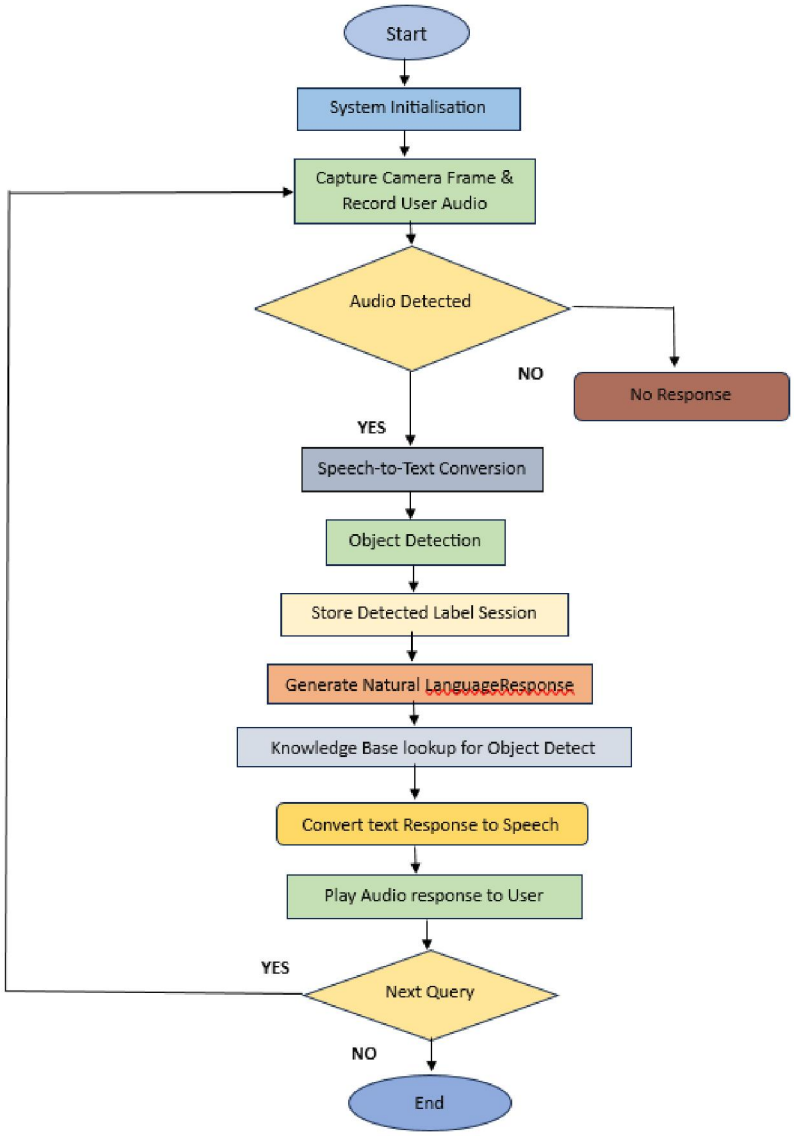

Fig.2 Camera interaction interface of VisionTalk

IV. SYSTEM IMPLEMENTATION

The VisionTalk system is implemented using modern web technologies and artificial intelligence frameworks to enable real-time visual interaction.

The VisionTalk interface allows users to interact with the system using a live camera feed and voice interaction. The interface captures images from the camera and sends them to the object detection module.

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 6, Issue 9, March 2026
Impact Factor: 8.2

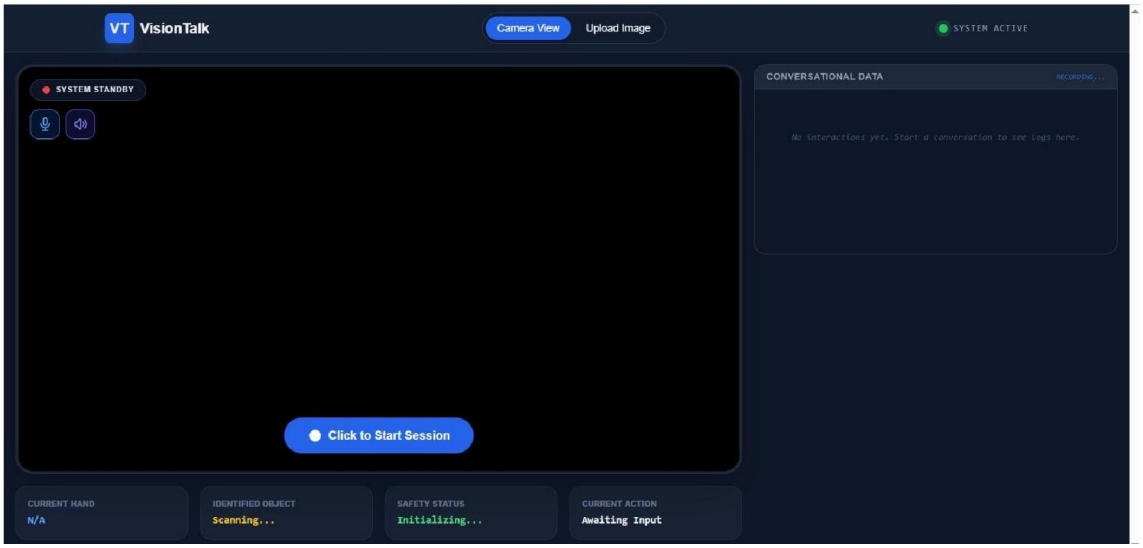

Fig. 3 Real-Time Camera Interaction Interface of VisionTalk

The front-end interface is developed using React and TypeScript. These technologies enable the system to capture real-time camera input from the user's device and display the AI-generated responses interactively.

The system captures video frames using browser APIs and converts them into image data. These frames are then processed by the AI models for object detection and contextual analysis.

Object detection is performed using YOLO-based models, which are capable of identifying multiple objects within an image. YOLO is chosen due to its high speed and suitability for real-time applications.

In cases where the object detection model cannot accurately identify an object, the system uses multimodal reasoning through Gemini AI. The Gemini model analyzes the visual content along with the user's voice input to generate contextual explanations. The system also allows users to upload images manually for analysis. The uploaded images are processed by the object detection model and multimodal AI reasoning module.

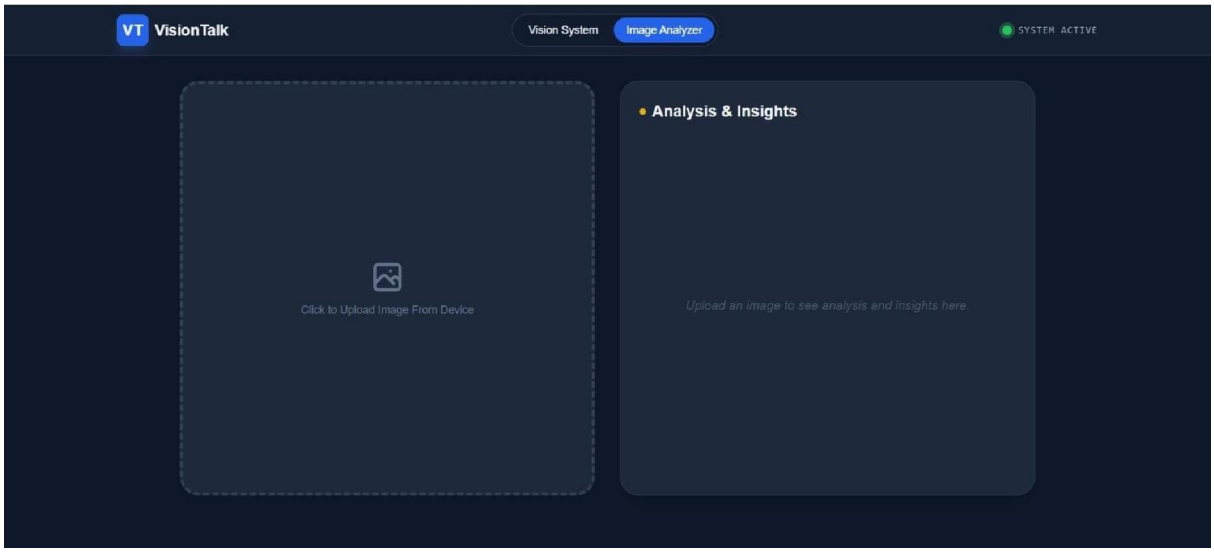

Fig. 4 Image Upload Interface of VisionTalk

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 6, Issue 9, March 2026 Impact Factor: 8.2

Speech interaction is implemented using audio processing techniques that convert text responses into spoken output. This allows users to receive guidance through voice responses, improving usability and accessibility.

The backend system is implemented using Python-based frameworks that manage AI model communication and image processing tasks.

Through the integration of these technologies, VisionTalk provides a real-time interactive AI system capable of identifying objects and providing contextual guidance to users.

V. RESULTS

The VisionTalk system was tested using various real-world objects to evaluate its effectiveness in object identification and user interaction. Experimental results show that the system can successfully detect common objects and generate contextual explanations through multimodal reasoning.

Fig. 5 shows an example of the system analyzing an uploaded image of an electric kettle and generating contextual guidance regarding its usage and maintenance.

When the object detection model identifies an object, the system displays the object label and provides guidance related to its usage. If the detection model fails to identify the object accurately, the multimodal AI model analyzes the visual input and generates a contextual explanation.

The VisionTalk system was evaluated using various real-world objects to analyze its object detection and interaction capabilities. Fig. 6 shows an example of real-time object detection performed by the system using the YOLOv8 model. The system successfully detects the object and displays the label along with interaction details such as the user's hand position, safety status, and current action.

The detected object in this example is an electric mosquito swatter. The system identifies the object and provides contextual information about its usage and safety considerations through the AI reasoning module.

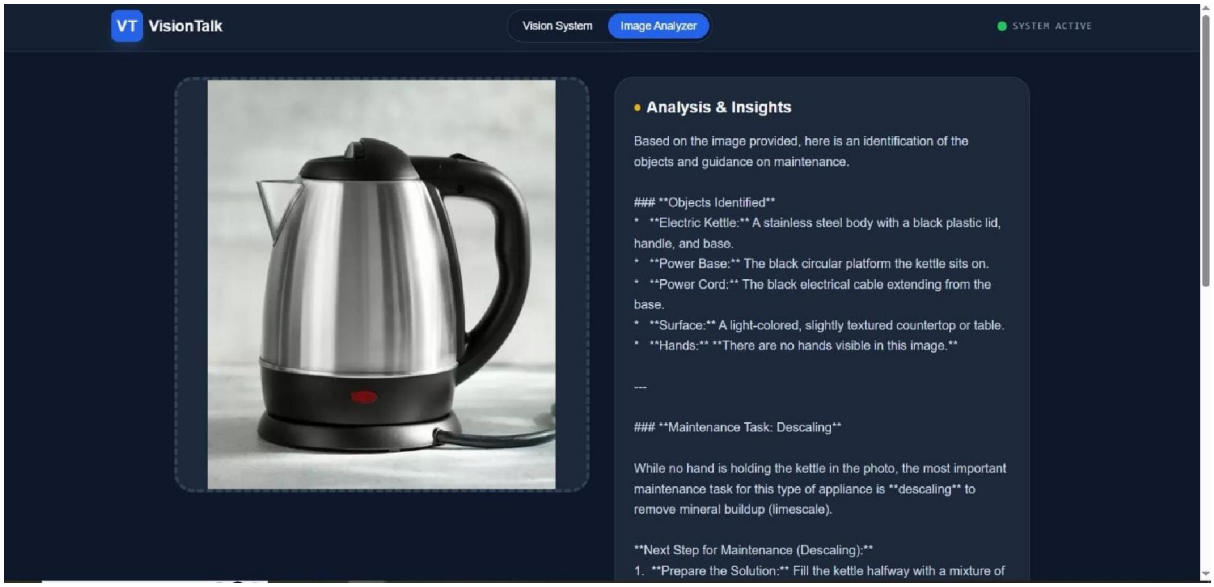

Fig. 5 AI-Based Object Analysis and Explanation Generated by VisionTalk

The speech interaction feature enables users to receive responses in audio format, which improves usability and accessibility.

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 6, Issue 9, March 2026
Impact Factor: 8.2

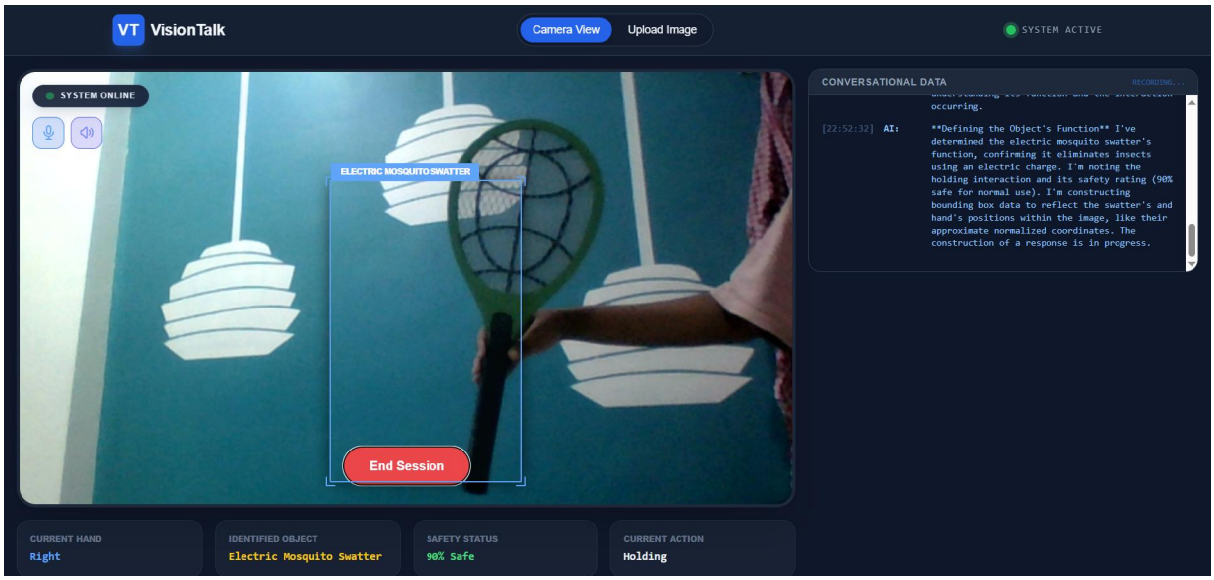

Fig. 6 Real-Time Object Detection and Interaction Interface of VisionTalk

Compared to traditional object detection systems, VisionTalk provides enhanced interaction capabilities by combining visual recognition with contextual reasoning and speech output.

VI. CONCLUSION

This paper presented VisionTalk, an AI-based visual understanding and speech interaction system designed to assist users through real-time object identification and contextual guidance. The system integrates computer vision, multimodal AI reasoning, and speech interaction to create an intelligent assistive platform.

The proposed system demonstrates the potential of combining object detection models with advanced multimodal AI systems to enhance human-computer interaction. By providing step-by-step guidance and speech feedback, VisionTalk can support users in performing tasks and understanding unfamiliar objects.

Future work may focus on improving detection accuracy, integrating cloud-based storage systems, and expanding the system to support mobile devices and augmented reality environments.

ACKNOWLEDGMENT

We would like to express our sincere gratitude to our project guide Siva Kumar Mondru, Assistant Professor, Department of Information Technology, Ace Engineering College, for his valuable guidance and continuous support throughout this research work.

We also thank the Department of Information Technology, Ace Engineering College, for providing the necessary resources and environment to carry out this study.

REFERENCES

[1] A. Ruan et al., "DetGPT: Detect What You Need via Reasoning," arXiv preprint arXiv:2305.14167, 2023.

[2] S. Antol et al., "VQA: Visual Question Answering," IEEE International Conference on Computer Vision (ICCV), 2015.

[3] P. Anderson et al., "Vision-Language Navigation: Interpreting Visually-Grounded Navigation Instructions," IEEE CVPR, 2018.

[4] J. Redmon et al., "YOLOv3: An Incremental Improvement," IEEE Conference on Computer Vision and Pattern Recognition, 2018.

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 6, Issue 9, March 2026
Impact Factor: 8.2

[5] S. Zhang et al., "A Novel Ultrathin Elevated Channel Low-Temperature Poly-Si TFT," IEEE Electron Device Letters, 1999.
[6] Google AI, "Gemini: A Multimodal AI Model," Google Research, 2024