VAYU-VISION ASSISTED YAAN FOR UTILITY

Atharva A Khetale[1], Chaitali Godse[2], Sangharsh Syal[3]

[1]Watumull Institute of Engineering and Technology, Ulhasnagar, India
[2]Watumull Institute of Engineering and Technology, Ulhasnagar, India
[3]Watumull Institute of Engineering and Technology, Ulhasnagar, India
atharvakhetle177@gmail.com, chaitali268@gmail.com , sangharshsyal77@gmail.com

Keywords: AI VISION, ESP32-CAM, SMART SURVEILLANCE, CHILD/PET MONITORING, AGENTIC AI.

Abstract

This paper presents VAYU (Vision Assisted Yaan for Utility), a novel autonomous home assistant and surveillance robot designed to integrate computer vision, artificial intelligence (AI), in a single mobile platform. VAYU combines low-cost hardware such as the ESP32-CAM with AI-enabled navigation, human/pet detection. Its multimodal agentic AI system enables VAYU to understand and react intelligently to its environment, providing security, convenience, and assistive functionality within a domestic setting. This system architecture and implementation offer a promising solution for next-generation home automation and monitoring system

1. Introduction

As smart home technologies continue to evolve, the demand for autonomous, intelligent, and multifunctional home assistants has surged. VAYU helps to address this demand by fusing hardware efficiency, artificial intelligence, and user-centric functionality into a compact, mobile platform. Its mission is to provide continuous surveillance and personalized assistance. Compared to existing smart devices and surveillance systems, VAYU introduces mobility, contextual understanding, and proactive decision-making, making it a unique and comprehensive home assistant.

This paper introduces a mobile robotic system that blends real-time computer vision, wireless control, and rugged mechanical design. Our Robot is constructed from a nylon framework and fitted with four heavy-duty wheels; it streams live video via ESP32-CAM to a remote AI server.
At the core of the VAYU's intelligence is the use of **LLaVA** (**L**arge **L**anguage and **V**ision **A**ssistant), an advanced vision-language model that can process video frames, interpret gestures or detect objects, and send actionable commands to the robot. [1]

1.1 Literature Survey

The idea of building an intelligent and interactive domestic robot, like VAYU is inspired by several important contributions across AI, vision, and human-robot interaction.

Verma [1] introduced the concept of LLaVA, which is a multimodal AI system. This system combines visual perception with language understanding. This model helps the robots in interpreting gestures and thus respond through contextual awareness an ability which VAYU adapts in real-time, for natural interaction.

In the field of embedded vision, Singh [11] showed how we can use OpenCV effectively in autonomous line-following robots. His work demonstrated that real-time computer vision could be deployed even on limited hardware, providing the groundwork for VAYU's perception system.

J.Kim [12] in his work took a step further and applied the ESP32-CAM and OpenCV for real time surveillance. The system used by them was static, but it proved that affordable, embedded setups could be used for 24/7 visual monitoring. This concept is used by VAYU for mobility and AI based scene interpretation.

Liu et al. [13] developed a more advanced version of LLaVA capable of understanding multimodal prompts and generating intelligent responses. Their work helped

us in developing systems like VAYU which can not only see but also understand and act in ways that resemble human-like reasoning.

Mehta [14] explored Gesture-controlled systems which used ultrasonic sensors and recognition based on images in an Arduino-powered robot. Even though these gestures were limited and predefined, they marked an important step in making robot control more intuitive. We explored this direction for VAYU and used dynamic, AI-powered gesture processing.

Finally, Ahuja [15] in his paper has discussed the growing need for privacy in home robotics. Their work which has emphasis on local data handling and user control, is central to VAYU's privacy-aware design. It ensures that sensitive data like voice and video are processed securely within the system.

Together, these studies have laid the foundation for VAYU's design which combines affordable hardware, advanced AI tools, an interactive control system, and a privacy-aware architecture all together in a unified robotic assistant.

2. System Architecture

The following is the basic design which includes the hardware and software specifications.

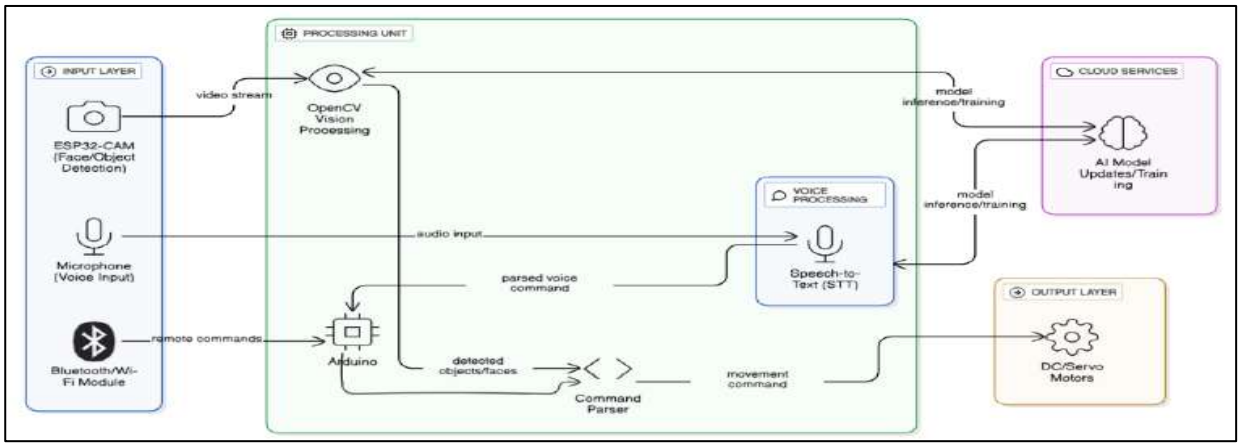

Fig 1: System Architecture

2.1 Hardware Specifications

The hardware used includes the ESP32-CAM Module, Arduino board, DC, and Servo motors.

2.1.1 ESP32-CAM Module:
ESP32-CAM Module is a low-cost, Wi-Fi-enabled camera module which is based on the OV2640 sensor. This sensor is responsible for live video capture and streaming. [2] This module has a microcontroller with built-in camera and Wi-Fi capabilities which captures video and streams it over Wi-Fi to a Python server. This allows the robot to send live video feed to the server so that it can do real-time processing and help in decision making.

Wi-Fi Protocol: This Protocol is used to establish communication between the ESP32-CAM and the Python server.
Video Streaming: The camera module also streams live footage for processing. The following are the specifications for the same.

- RAM: 520KB SRAM
- Flash Memory: 4MB
- Power Supply: 5V via USB or external supply

2.1.2 Arduino Board:
Arduino UNO is a microcontroller board that is based on the ATmega328P. It has 14 digital input/output pins of which 6 can be used as PWM outputs, 6 are analog input pins. It also has a 16 MHz ceramic resonator, a USB connection, a power jack, an ICSP header and a reset button.[3]

The Arduino board acts as the controller for DC motors and servo motors. It receives serial commands from the ESP32 and processes these commands to drive the physical movement of the robot.

The specifications are as follows.

- Microcontroller: ATmega328P
- RAM: 2KB (Arduino Uno)
- Flash Memory: 32KB (Arduino Uno)

2.1.3 Motor Driver Shield:
The Motor Driver Shied provides an easy and convenient way to control motors, and it integrates

directly with the Arduino Board. This shield handles the motor control, its direction, along with its speed without requiring separate motor driver modules. [4]. The key features include a Plug-and-play integration with the Arduino Board for motor control. It can help to drive DC motors, servo motors, and stepper motors. The Motor Driver Shield is often equipped with built-in motor protection and current sensing.[5]. The Specifications are as follows:

- The Shield is compatible with Arduino Uno and other similar boards.
- The Voltage range is typically from 5V to 12V depending on the motors.
- It provides PWM for motor speed control.

2.1.4 DC and Servo Motors

DC Motors and Servo Motors are used for the robot's mobility and articulation. The motors are responsible for basic movements (forward, backward), while servos can control more precise movements such as tilting or rotating components of the robot.

2.2 Software Specifications

The software which is used includes Arduino IDE, Python 3.x environment and LLaVA model.

2.2.1 ESP32 Firmware

The ESP32 Firmware is used create a custom code that initializes the camera, streams the MJPEG video over HTTP, and also communicates with the main processor. [6]

2.2.2 Python Server

We have used the Python Server as the backend for video processing; wherein real-time frames received from the ESP32-CAM are parsed and analyzed for further processing

OpenCV: OpenCV is used for image processing and computer vision. This library helps us in detection of objects, recognizing gestures, and other visual inputs that can trigger specific actions in our robot.[7]

2.2.3: LLaVA (Large Language and Vision Assistant)

LLaVA is an end-to-end trained large multimodal model. This model is designed to understand and generate content which is based on visual inputs (images) as well as textual instructions. It also combines the capabilities of a visual encoder and a language model so that it can process and respond to multimodal inputs.[8][9]

The LLaVA model is used for advanced AI processing, in which the system interprets commands or inputs from the camera feed and transforms these commands into parsed commands. This model is also used to interpret visual information and thus provide a contextual understanding which helps to drive the robot's actions based on its surroundings.

Parsed Commands (AI Decision Logic): The Parsed Commands are generated after processing the video stream and identifying key actions (e.g., object detection, gestures, etc.). These commands define the specific actions that need to be executed by the robot.

3. Design and Methodology

Our prototype is designed with a modular and robust architecture that supports mobility, sensing, and interaction. As shown in Figure 2, the robot's structure is built by using a durable nylon box frame, which offers a lightweight yet rigid enclosure that houses all critical electronic components and also provides stability during movement.

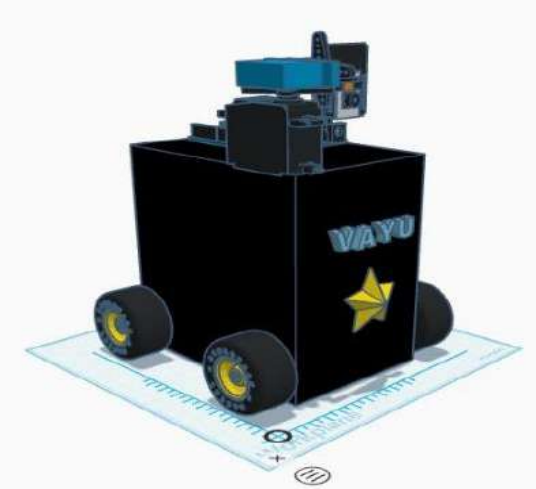

Fig 2: CAD render of the VAYU robot showing external chassis and internal components layout

A wheel and servo motor system is used to achieve VAYU's mobility, which allows it to perform basic manoeuvres such as forward motion, turning, and pivoting. The mechanical setup is kept simple enough to allow seamless integration making sure that the movement is not impacted.

Internally, the robot includes a microcontroller-based control unit that supports key features like:

- Obstacle detection via ultrasonic or infrared sensors
- Voice interaction through speech recognition modules
- Gesture response, which is processed via an external server and then interpreted using computer vision

This modular layout helps to simplify debugging and future upgrades. It makes the system adaptable for applications in human-robot interaction, surveillance, and autonomous navigation tasks.

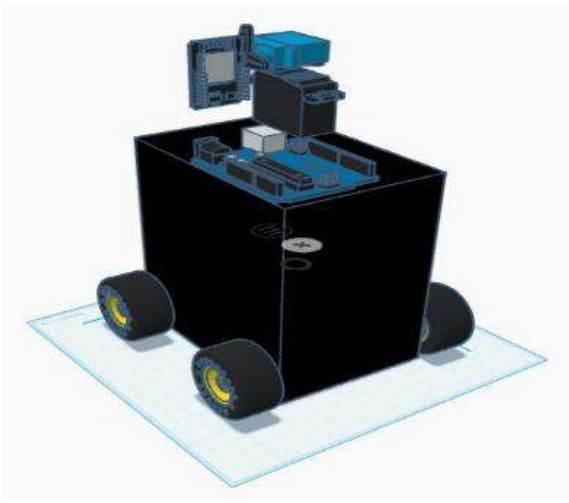

Fig 3: Rear view of the VAYU robot showing internal hardware layout including microcontroller and sensor modules.

The system integrates a real-time gesture and voice controlled robotic platform that utilizes a combination of edge computing and AI-based vision models. The complete workflow is illustrated in Figure 1. This workflow outlines the sequential processing starting from video capture to robotic actuation.

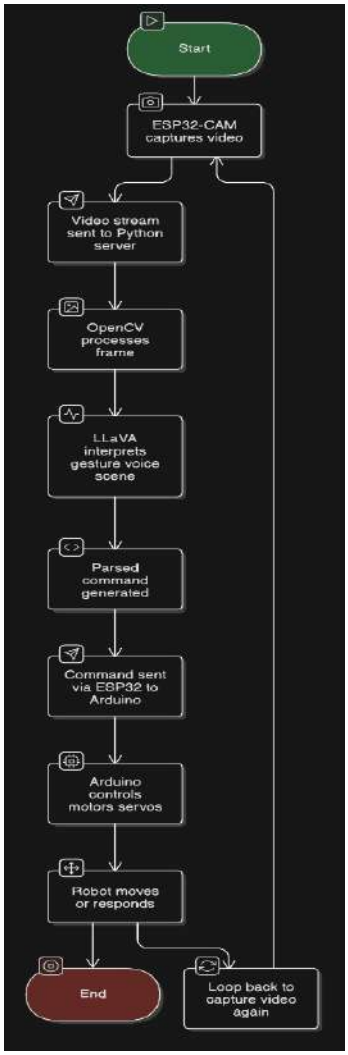

Fig 4: System workflow for gesture/voice-controlled robotic interaction using ESP32-CAM, OpenCV, LLaVA, and Arduino

The control logic of the proposed VAYU robot is as shown in Figure 4. The system employs an ESP32-CAM for visual input, OpenCV for frame analysis, and LLaVA (Large Language and Vision Assistant) for gesture and voice interpretation. Commands are parsed and executed through an Arduino-based actuator control mechanism.[1][8][9]

3.1 System Workflow

The workflow is outlined below to complement Fig. 4

- Start:
  The system will power on and initialize all the modules.
- ESP32-CAM captures video:
  The ESP32-CAM module begins to stream live video of the surrounding environment.

- Video stream sent to Python server: The video stream is transmitted to a Python-based processing server in real-time
- OpenCV to processes frame: OpenCV is used to process frames from the video stream to extract visual features.
- LLaVA interprets gesture/voice scene: The processed frame is analysed by the LLaVA model to detect gestures or voice-based cues
- Parsed commands are generated: A structured command is generated based on the interpreted scene.
- Command is sent via ESP32 to Arduino: This structured command is sent through the ESP32 module to the Arduino microcontroller.
- Arduino controls motor servos: The Arduino now actuates the corresponding motors based on the instructions received.
- Robot moves or responds: The robot executes the intended action, such as moving forward, turning, or stopping.
- Loop back to capture video again: The system continuously loops back to capture new video, which helps to enable real-time operation.
- End: The cycle continues until it is manually stopped.

4. Functional Modes

The VAYU robotic system is designed to operate in multiple functional modes, each catering to distinct real-world applications. The system adapts dynamically based on user interaction, environment, and predefined tasks. The primary operational modes are as follows

4.1 Surveillance Mode

In this mode, the robot autonomously patrols predefined routes within indoor environments. It utilizes onboard sensors for motion and sound detection and provides

- Real-time alerts upon detecting unexpected activity.
- Continuous logging of surveillance data for later review.

4.2 Child and Pet Monitoring

This mode allows the robot to follow and monitor specific individuals, such as a child or a pet. Using image and audio analysis, the robot

- Captures key moments and behaviours.
- Identifies hazardous situations (e.g., proximity to stairs or sharp objects).
- Detects anomalies such as no activity for a long time or distress.

4.3 Assistant Mode

In Assistant Mode, the robot serves as a mobile companion that can perform interactive tasks. These tasks include the following

- Voice based interactions for giving reminders, updating schedules, and responding to queries.
- Mobile communication support for interaction within the family with the help of updates and alerts.

5. Artificial Intelligence Design

The VAYU robot incorporates a layered AI architecture that enables autonomous decision making through sensor fusion, machine perception, and context aware behaviour generation.

5.1 Multimodal AI System

The robot utilizes a multimodal approach that combines visual, auditory, and motion based data streams. This fusion helps in adaptive interaction with the environment. Local edge devices are used to process routine tasks in real-time.

5.2 Computer Vision Techniques

Computer vision forms the core of the robot's perception system. The system uses models trained on object and human recognition datasets to

- Detect and identify persons or objects in the environment.
- Analyse scenes to determine activity context and user behaviour.

5.3 Navigation and Mapping

For navigation and mapping the robot uses OpenCV and LLaVA model for the following

- Obstacle detection and avoidance.

- Path planning to follow patrol routes or return to docking stations.

6. Implementation Details

The physical and software infrastructure that supports VAYU ensures reliable, secure, and efficient operation of the Robot in residential environments.

6.1 Network Architecture

The system is configured such that it can operate on standard Wi-Fi infrastructure and it can support both Access Point (AP) and Station (STA) modes. The ESP32-CAM can stream video over the root stream endpoint (/) on port 81. Commands and data are exchanged over HTTP and MQTT protocols for low latency performance.[10]

6.2 Security and Privacy

To address Privacy and data protection the following is done:

- There is End-to-end encryption of video streams and communication channels.
- Local processing of sensitive inputs is done such as voice and visual recognition to prevent cloud leakage.
- The privacy zones are User-defined that block or blur sensitive areas in the video feed.
- The recording settings can be configured and controlled via a secure web interface.

7. Results

We successfully tested our developed system across multiple scenarios which involved gesture recognition, voice interaction, and autonomous robotic control.

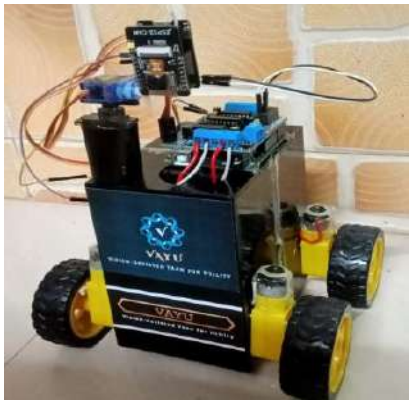

Fig 5: Assembled VAYU Robot with ESP32-CAM and Arduino Control Unit

Figure 5 shows the assembled VAYU robot, with the hardware integration of ESP32-CAM, Arduino Board, and motor driver components. Our system successfully executed commands parsed from gesture and voice input, by validating the complete pipeline from image acquisition and AI interpretation to motor actuation.

Figure 6 illustrates the Python based interactive vision module that interprets live camera input from the ESP32-CAM. The AI model identifies and describes visual scenes in natural language, accurately detects human presence, their attire, and the background details in real time. This demonstrates the system's capacity for multimodal scene understanding.

This real-time stream was also processed through OpenCV and LLaVA modules with minimal latency, to enable continuous user interaction. The robot exhibited responsive behavior in navigation and monitoring tasks that confirmed the feasibility of deploying VAYU for human/pet tracking, surveillance, and basic home automation functions.

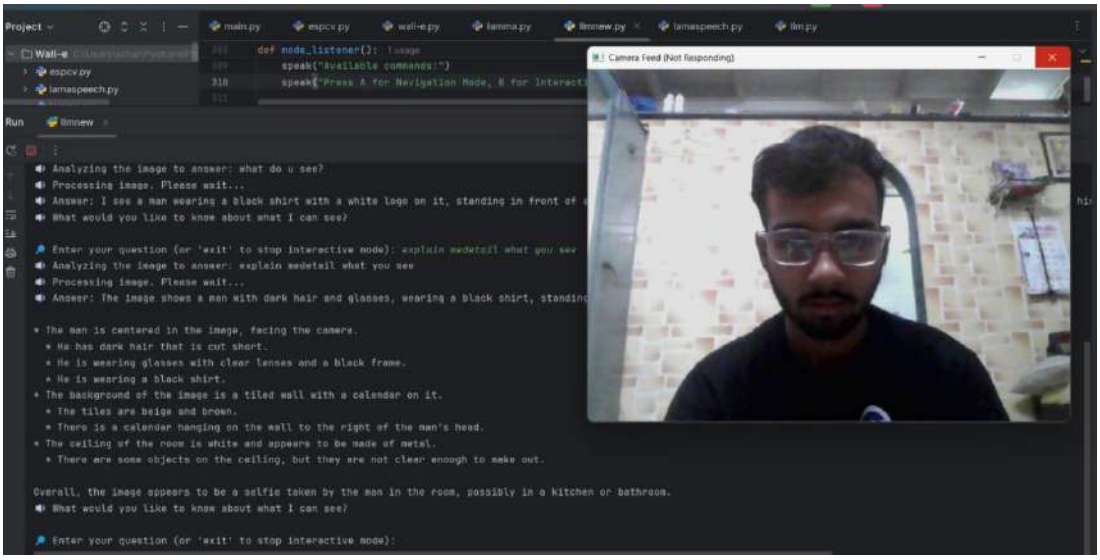

Fig 6: Interactive AI Vision Output [Human Detection]

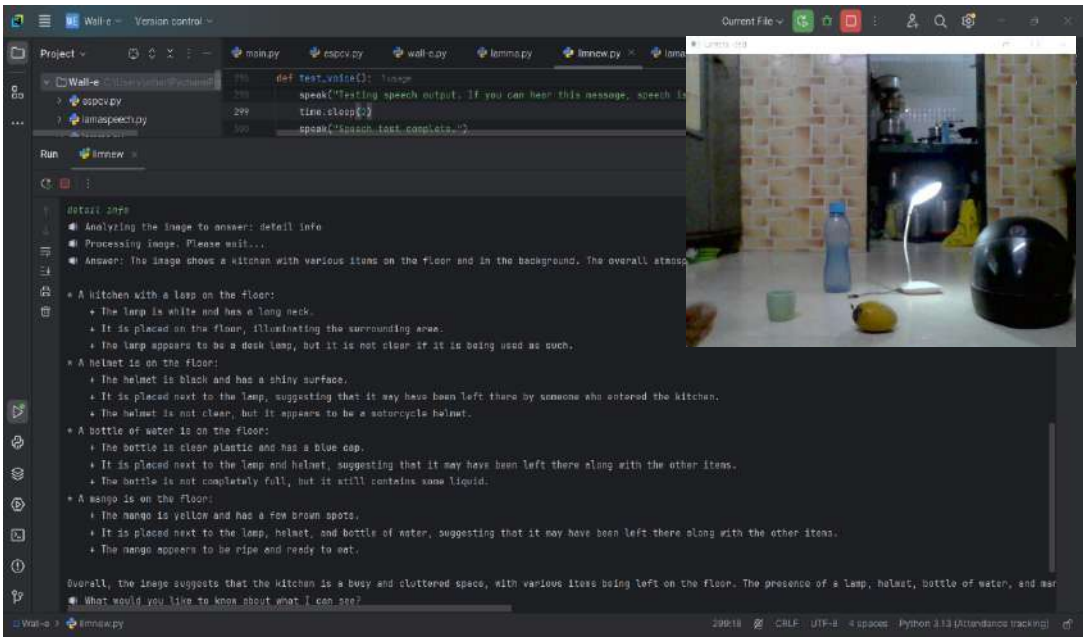

Fig 7: Interactive AI Vision Output [Object Detection]

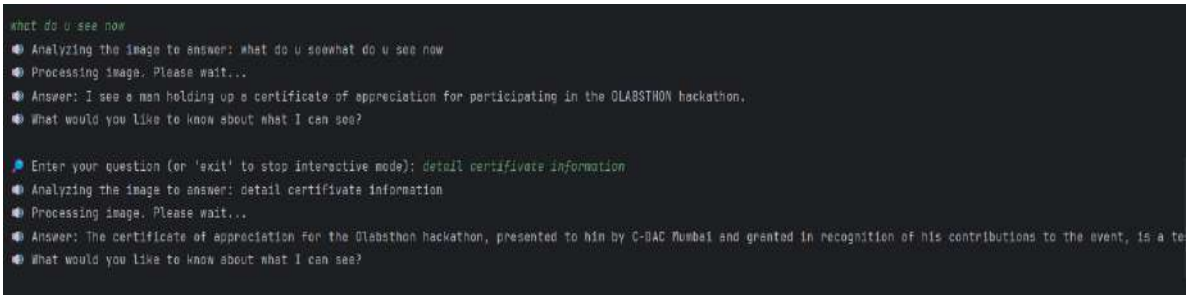

Fig 8: *Interactive AI Vision Output [Detailed Description]*

8. Performance Evaluation and Metrics

We conducted several tests to validate the effectiveness of the VAYU system for real-world conditions. These test focused on different operational modules. The table below gives a summary of the quantitative results obtained during field trials.

8.1 Module-Wise Performance Metrics

Feature / Task	Test Cases	Successful Attempts	Failures	Success Rate
Face/Object Detection (via ESP32-CAM)	50	47	3	94%
Voice Command Recognition (Offline TTS/STT)	100	92	8	92%
Real-Time Signaling via Bluetooth/ Wi-Fi	30	28	2	93.3%
Navigation to Predefined Target Locations	40	36	4	90%
Video Streaming Quality (ESP32-CAM)	10 mins avg	8 mins smooth, 2 mins lag	—	80%
Battery Endurance (on full charge)	—	1.5–2 hours average	—	—
Responsiveness to Obstacle Avoidance	25	21	4	84%

8.2 System Latency Metrics

Component	Avg. Response Time
Voice Command to Action	~850 ms
Camera Detection Frame Rate	~15 FPS
TFT Display Refresh Time	~200 ms
Video Stream Delay (avg)	~1.3 sec

8.3 Environmental Testing Results

Lighting Condition	Detection Success Rate
Bright Indoor Light	96%
Natural Daylight	92%
Low Light / Evening	79%
Near Darkness (IR only)	51%

8.4 Reliability over Time

Mean Time Between Failures (MTBF):	~3.5 hours
Average Reboot Recovery Time:	~12 seconds
System Downtime Per Day (during test):	<5%

9. Conclusion

VAYU isn't just a robot but it's a step to make our homes more intelligent, responsive, and secure. VAYU brings together Artificial Intelligence, vision, and mobility in a small package. VAYU's ability to adapt so that it can learn, respond, and assist makes it a unique robot.

VAYU delivers an interactive experience through real time gesture and voice recognition, which goes beyond basic automation. It has a modular build which helps the robot to evolve, and open doors for applications related to elder care, smart surveillance, and even companionship.

In the future, our aim is to enhance VAYU's emotional intelligence, extend its outdoor functionality, and also make it more energy-efficient. VAYU is not just a project but it's also our vision of what the next generation of assistant robots can be: helpful, aware, and truly human centred.

10. Acknowledgments

We would like to thank our management, and faculty of Watumull Institute of Engineering and Technology, Ulhasnagar, who provided us with the support that was needed for this project.

We extend our sincere gratitude to Mr. Arvind M. Khetale for his help in the structural design of VAYU. His expertise and technical support has helped us to successfully develop the project's framework.

We would also like to extend our appreciation to our peers and colleagues who have provided constructive feedback and kept us motivated.

11. References

[1] U. Verma, "Introduction to LLaVA: A Multimodal AI Model,"*Medium*,Dec.27,2023. https://medium.com/@ud.uddeshya16/introduction-to-llava-a-multimodal-ai-model-2a2fa530ace4

[2] Ai-Thinker, “ESP32-CAM WiFi + Bluetooth Camera Module,” *Ai-Thinker*, 2023. [Online]. Available: https://loboris.eu/ESP32/ESP32CAM%20Product%20Specification.pdf

[3] Arduino, "Arduino Uno Rev3," *Arduino Documentation*. https://docs.arduino.cc/hardware/uno-rev

[4] Arduino, “Arduino Motor Shield Rev3,” *Arduino Official Store*,2023. https://store-usa.arduino.cc/products/arduino-motor-shield-rev

[5] Elecrow, “Arduino Motor/Stepper/Servo Shield,” *Elecrow Wiki*,2023 https://elecrow.com/wiki/arduino-motorstepperservo-shield.html

[6] Espressif Systems, "ESP-AT Binaries for ESP32," *Espressif Documentation* https://docs.espressif.com/projects/esp-at/en/latest/esp32/AT_Binary_Lists/esp_at_binaries.htm

[7] OpenCV, “OpenCV-Python Tutorials,” *OpenCV Documentation*, https://docs.opencv.org/4.x/d6/d00/tutorial_pyroot.html

[8] LLaVA Team, "LLaVA: Visual Instruction Tuning," *LLaVA Official Website*, https://llava-vl.github.io/

[9] LLaVA Team, “LLaVA: Large Language and Vision Assistant,” *GitHub Repository*, https://github.com/haotian-liu/LLaVA

[10] R. Santos, “ESP32-CAM: Set Access Point (AP) for Web Server (Arduino IDE),” *Random Nerd Tutorials*. https://randomnerdtutorials.com/esp32-cam-access-point-ap-web-server/

[11] R. Singh, “Using OpenCV for Autonomous Line-Following Robots,” *International Journal of Embedded Systems*, vol. 8, no. 3, 2021.

[12] J. Kim, “Real-Time Surveillance with ESP32-CAM: A Cost-Effective Solution,” *Journal of IoT Innovations*, vol. 9, 2022.

[13] H. Liu et al., “LLaVA: Visual Instruction Tuning,” arXiv preprint arXiv:2304.08485, 2023.

[14] D. Mehta, “Gesture Controlled Arduino Robot using Ultrasonic and IR Sensors,” *IEEE Conference on Human-Robot Interfaces*, 2020.

[15] P. Ahuja, “Privacy-Preserving Robotics for Smart Homes,” *Smart Systems Journal*, vol. 5, 2021.

12. Bibliography of the Authors

Atharva Khetale

Third-Year Computer Engineering student at Watumull Institute of Engineering and Technology. His interests include robotics, embedded AI, and cloud-connected IoT systems. He led the hardware integration and AI pipeline development of the VAYU project.

atharvakhetle177@gmail.com

Chaitali Godse

Assistant Professor at Watumull Institute of Engineering and Technology. She specializes in computer vision. She guided the conceptual design and supervised the implementation phases of the VAYU project.

chaitali268@gmail.com

Sangharsh Syal

Associate Professor at Watumull Institute of Engineering and Technology, with expertise in wireless networks, and control systems. She mentored the team with AI based assistance and system optimization.

sangharshsyal77@gmail.com