AIris | AI-Powered Assistant for the Visually Impaired

ADIB AR RAHMAN KHAN
SAUMIK SAHA KABBYA

SUPERVISOR
DR. MOHAMMAD RASHEDUR RAHMAN

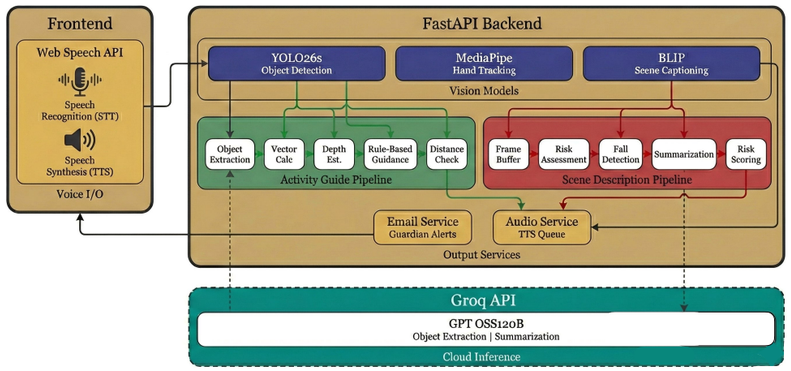

Figure 1. Complete system architecture of AIris showing three-tiered design: Frontend, FastAPI Backend (vision models, Activity Guide pipeline, Scene Description pipeline, and output services), and Groq API cloud inference. The diagram illustrates data flow between components, with Activity Guide (green) and Scene Description (red) pipelines clearly separated.

Abstract

Visually impaired individuals face significant **challenges** in independently **navigating** their environment and **locating** objects. Existing assistive technologies often suffer from high latency, reliance on unstable cloud connections, and a passive approach that describes scenes without offering actionable guidance. **AIris** addresses these limitations by introducing a r**eal-time**, **server-based AI assistant** designed for **active interaction**. By integrating state-of-the-art computer vision models (**YOLO26s** for object detection, **MediaPipe** for hand tracking, and **BLIP** for scene analysis) with a robust **LLM reasoning engine** (**GPT OSS120B via Groq API** for object extraction and summarization), AIris provides two core modes: **Active Guidance**, which directs the user's hand to specific objects with precise audio cues using a **novel rule-based geometric algorithm** (**no LLM** required for guidance generation), and **Scene Description**, which offers **continuous** environmental **awareness** with advanced fall detection and **automated** guardian email alerts. The system features complete handsfree voice control via **Web Speech API** (native browser **STT**/**TTS**), prioritizes privacy through local-first processing, and achieves **sub-2-second** response times via low-latency cloud inference, making **real-time** assistance a practical reality.

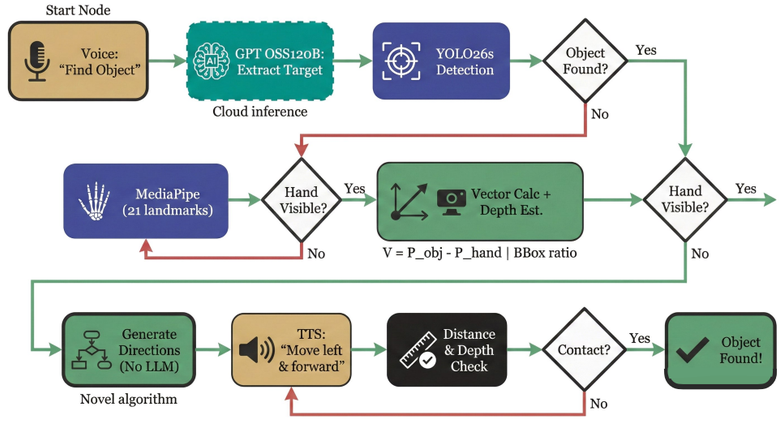

Figure 2. Active Guidance mode decision loop flowchart. The system processes voice commands through GPT OSS120B for object extraction, performs YOLO26s object detection and MediaPipe hand tracking, calculates spatial vectors and depth estimates, and generates real-time directional guidance using a novel rule-based algorithm (no LLM). The loop continues until hand-object contact is confirmed, providing step-by-step audio instructions to guide users to target objects.

Novelty

Our system serves **two novel** operations. First, the **Active Guidance Mode (Targeted):** When a user asks to **find** an object (e.g., "Find my water bottle"), the system uses **GPT-OSS 120B (Groq API)** to extract the target object from the natural language goal. The system then activates **YOLO26s** (Ultralytics) to locate the target in real-time with high precision. Simultaneously, **MediaPipe** tracks the user's hand coordinates using **21 landmark points**. A proprietary geometric algorithm **calculates** the **3D spatial vector** between the hand centroid and object bounding box center, and estimates **relative depth** using **bounding box size ratios**. A **novel** rule-based **algorithm** (**no LLM**) **converts** this **vector** and **depth** information into **natural** language directional commands, generating **precise** audio **instructions** ("Move slightly right and forward") in **real-time** until hand-object contact is confirmed through distance and depth threshold analysis.

Complimenting this, the **Scene Description Mode (Continuous)**: For general awareness, the system employs **BLIP** (Bootstrapping Language-Image Pre-training) to capture **dense** semantic **captions** of the video feed at 2 FPS. These captions undergo **quick** keyword-based **risk assessment**, static **frame detection** (for **fall indicators**), and **transition pattern analysis**. Frame descriptions are accumulated in a buffer (5-10 frames) and processed by **GPT OSS120B** on the **Groq fast inference** engine to generate **contextual summaries**, calculate multi-factor **risk scores**, and **detect potential falls** through multi-method analysis. The system includes **automated** guardian **email** alerts with **configurable risk thresholds** (0.1-0.5) and provides daily and weekly activity summaries.

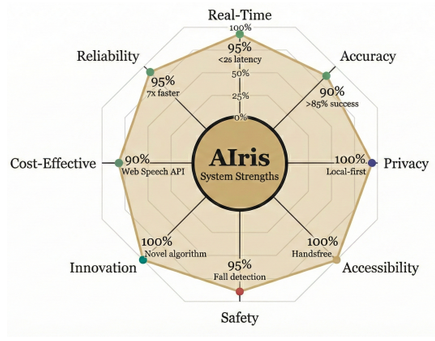

Figure 3. Octagonal radar chart displaying AIris system strengths across eight key dimensions: Real-Time Performance (95%, <2s latency), Accuracy (90%, >85% success), Privacy (100%, local-first), Accessibility (100%, handsfree), Safety (95%, fall detection), Innovation (100%, novel algorithm), Cost-Effectiveness (90%, Web Speech API), and Reliability (95%, 7x faster). The chart demonstrates comprehensive system capabilities with near-maximum performance across critical accessibility metrics.

Scalability

AIris is **architected** for **global scale**, designed to serve **millions** while maintaining transformative **personal impact**. We plan on implementing a **cloud-based SaaS infrastructure** that enables exponential growth **without** compromising real-time performance. To be built on **scalable VPS** infrastructure, **AIris** will serve from **hundreds** to hundreds of **thousands** simultaneously. Our **API-based architecture** will help to ensures consistent **sub-2-second** response times **globally**. Users can use **any device** with a camera and speakers/microphone to log into the system and gain assistance.

As for the **Market Reach,** with **2.5 million** visually impaired in Bangladesh (500K addressable) and **285 million** globally, **AIris** addresses **critical** needs for **independence** and **safety**. Our affordable pricing (*$9.99/BDT1000* a month) ensures accessibility **without compromising** sustainability.

Every subscription enables **algorithm improvements** and feature **expansion**. We are building a movement toward **universal accessibility** where technology **serves humanity**, **independence** is a right, and **AI** becomes a **force** for genuine **social transformation**.

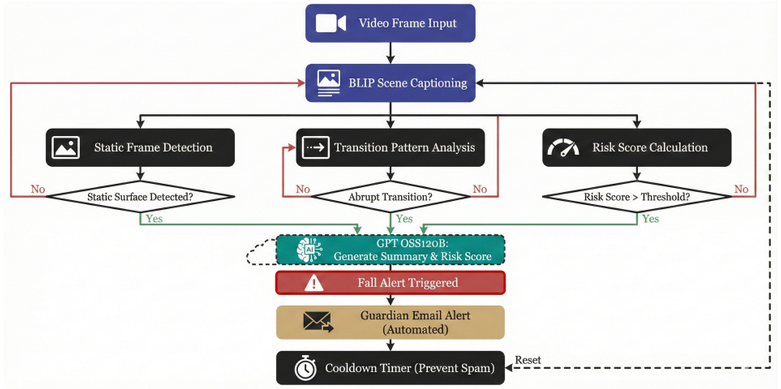

Figure 4. Multi-method fall detection algorithm flowchart. Three parallel analysis paths process BLIP-generated captions. All paths converge to GPT OSS120B for summary generation and risk scoring. When risk exceeds threshold, the system triggers fall alerts, sends automated guardian email notifications.

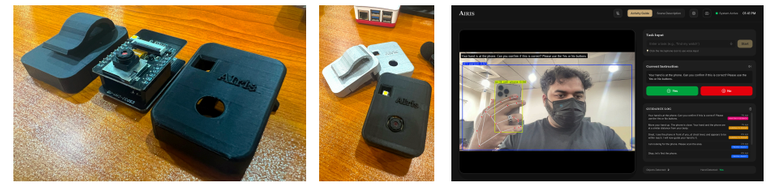

Figure 6.(A, B) Custom-designed hardware components: 3D-printed protective casing with "AIRIS" branding, ESP32-CAM-MB module, and mounting clip. **(C)** Real-time software showing object detection, MediaPipe hand tracking with 21-landmark skeletal overlay, and interactive guidance providing step-by-step instructions for object interaction. The interface demonstrates handsfree operation with voice input and audio capabilities.

Conclusion

AIris demonstrates that **complex**, **multi-modal AI** systems can be effectively engineered for **real-time** assistive applications on **consumer hardware**. By strictly **prioritizing latency** and **active guidance** over passive description, we have created a system that is **not** just an **observer**, but an **active participant** in the user's world. Our **novel** rule-based **guidance algorithm,** operating **without any LLM API** dependency for real-time performance, represents a **breakthrough** in assistive technology, achieving **sub-2-second** response times through **intelligent geometric vector calculation** and **depth estimation**. The integration of YOLO26s, MediaPipe, **proprietary fall detection**, and Web Speech API creates a comprehensive **solution** that **transforms** passive scene description into **actionable** assistance. This implementation sets a new **standard**, proving that cutting-edge AI can be simultaneously **powerful**, **accessible**, and **affordable**, a blueprint for the **future** of assistive technology that will continue to evolve as we **scale** from **hundreds** to **millions** of users, moving us **closer** to a world where technology truly "**opens eyes**".

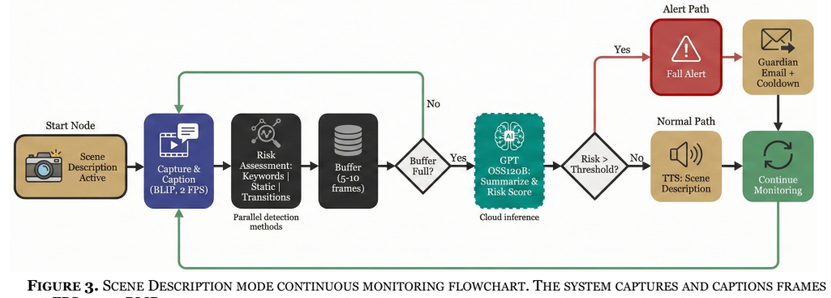

Figure 3. Scene Description mode continuous monitoring flowchart. The system captures and captions frames at 2 FPS using BLIP, performs multi-method risk assessment, accumulates frames in a buffer, and processes them through an LLM for summarization and risk scoring. The system branches into normal operation or alert path based on risk threshold, then loops back to continuous monitoring.