Team Name: CappyBara

Team Lead: Aniket Baravkar

1. Problem Statement

Visually impaired individuals face daily challenges in navigating unfamiliar environments, detecting hazards, and interpreting complex social cues. Traditional aids like walking canes or basic voice assistants are limited to obstacle detection or rigid voice commands, offering minimal situational understanding. Without contextual awareness, independent movement becomes difficult and sometimes dangerous. Our solution empowers the blind with real-time visual narration, contextual scene understanding, and intelligent navigation support using Generative AI. The goal is to enable them not only to "see" but to **comprehend** their surroundings, promoting confidence, independence, and inclusion in both public and private spaces.

2. Target Audience & Context

The application is designed for blind individuals living in urban and semi-urban environments where mobility and environmental complexity are high. Whether navigating busy intersections, using public transport, or simply identifying people and objects around them, this group needs more than basic obstacle avoidance—they require a companion that can **interpret** their environment. The application adapts to indoor and outdoor contexts and addresses a key accessibility gap by combining **real-time intelligence, offline functionality**, and a conversational interface for ease of use.

3. Use of Gen-AI

Generative AI is central to our solution. A fine-tuned multimodal model (based on BLIP-2 or LLaVA, trained on accessibility datasets like VizWiz) interprets live camera feeds to generate meaningful scene descriptions in natural language. It understands context, relationships between objects, and even social cues—for instance, detecting a person waving or an approaching vehicle. The system also supports conversational queries such as "Is it safe to cross?" or "Where am I?" Offline capabilities are ensured through a distilled version of the model, using MobileViT or TinyBLIP architectures, which still retain scene understanding without internet. The AI adapts outputs for clarity, emotional tone, and urgency—

ensuring users receive not just information, but insight conveyed in a way that's calming, understandable, and immediately actionable. Providing a **personalized, human-like assistant experience** for users with accessibility needs.

4. Solution Framework

Our system comprises a hybrid on-device and cloud-based architecture optimized for speed, safety, and context awareness. The key components include:

- **Camera Input + Preprocessing**: Captures and enhances real-time visual data using edge processing (OpenCV or MediaPipe).
- **Generative Scene Understanding (Gen-AI)**:
    - Cloud: Full model (BLIP-2, GPT-Vision) for rich scene analysis, object detection, spatial awareness.
    - Edge: Distilled lightweight model (TinyBLIP or MobileSAM) for offline inference.
- **Voice Interaction Layer**:
    - Uses Whisper ASR for voice input and Coqui TTS or Polly for lifelike narration.
    - Accepts natural language commands like "Tell me what's in front of me."
- **Navigation Engine**:
    - GPS + Maps API for real-time routing.
    - Indoor navigation supported via BLE beacons or Wi-Fi mapping.
    - Haptic and voice prompts for upcoming turns or alerts.
- **Safety Advisor**:
    - Detects traffic movement, pedestrian crossings, obstacles, and hazards like open manholes.
    - Uses time-of-day and environment detection (e.g., rain, dim light) to modify alerts dynamically.
- **Smart Mode Switching**:

    - Automatically switches between cloud and offline mode depending on connectivity.
- **Learning Layer**:
    - Learns user behaviour over time—preferred routes, walking pace, frequently visited areas—for personalized feedback.

All outputs are tailored for clarity, urgency, and emotional tone to reduce anxiety and enhance usability.

5. Feasibility & Execution

We will fine-tune open-source Gen-AI models (like BLIP-2 or LLaVA) using accessibility datasets (e.g., VizWiz, COCO with captioning). The MVP will be built on Android, leveraging TensorFlow Lite for model deployment and Google Maps API for navigation. Offline inference will be enabled through model quantization and pruning. Audio input/output will be handled using Whisper and Coqui TTS. The prototype can be developed within 4–6 months using a team of ML engineers, accessibility testers, and mobile developers. Accessibility standards like WCAG and screen-reader compliance will be prioritized.

6. Scalability & Impact

The solution is built to scale both vertically and horizontally. Technically, new languages, regional map data, and scene contexts can be added through modular training. Hardware scalability allows integration into smart glasses, AR devices, or IoT accessories. Partnerships with disability organizations, city transport authorities, and NGOs will help expand reach. With growing awareness and regulations around AI for accessibility, our solution has the potential to become a **standard assistive tool**, improving the quality of life for millions of users across geographies.

7. Conclusion / Bonus (Minimum Lovable Product)

We have successfully developed our **Minimum Viable Product** (MVP) and are excited to showcase it and build upon it further. The application empowers the blind community using generative AI for real-time scene narration, safe navigation, and hazard detection. With features like offline Gen-AI fallback, crosswalk alerts, and natural voice interaction, our solution demonstrates strong technical feasibility and commercial potential while addressing a critical accessibility gap.