Design and Development of WalkSense: A Vision-Language Assistant For Real-Time Scene Understanding For Visually Impaired Individuals

A Mid-Semester Progress Report for AIMLCZG628T: DISSERTATION

Student:
ANIKET AYODHYA CHAUHAN (2023AD05122)

Supervisor:
Manish Kumar (Ignisnova Robotics Private Limited, Navi Mumbai)

Enhance Environmental Awareness for Visually Impaired Individuals Through Real-Time, Multimodal AI

Core Objective

To design and implement a real-time assistive system that perceives real-world scenes through a camera and conveys contextual information via natural audio feedback.

Scope

The system is designed for local, on-device execution on a laptop or edge device, prioritizing low latency, reliability, and usability.

System Integrates

Computer Vision

For immediate object and hazard detection.

Vision-Language Models (VLM)

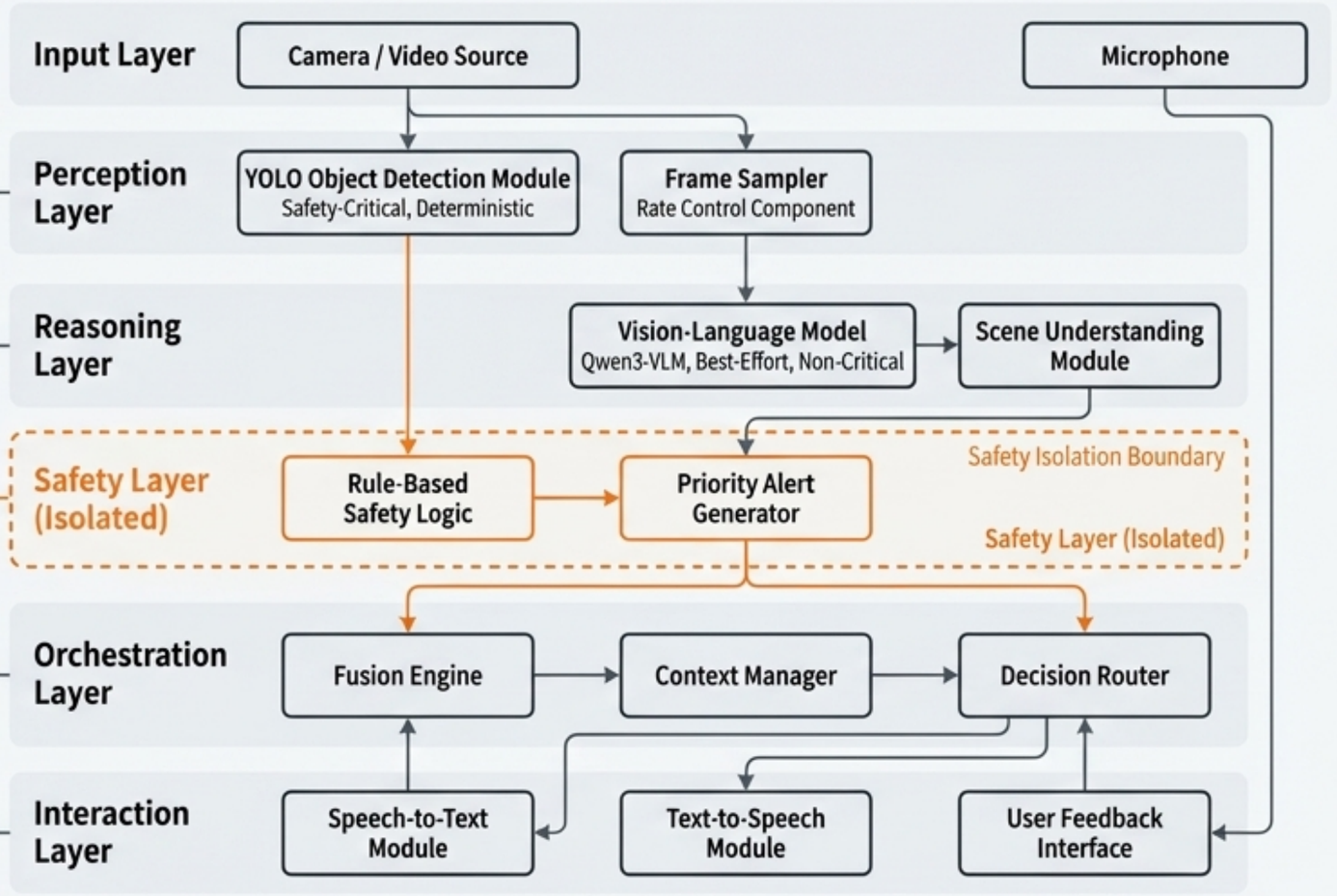
For deep, multimodal scene understanding and reasoning.

Speech Processing

For natural, hands-free voice interaction (STT/TTS).

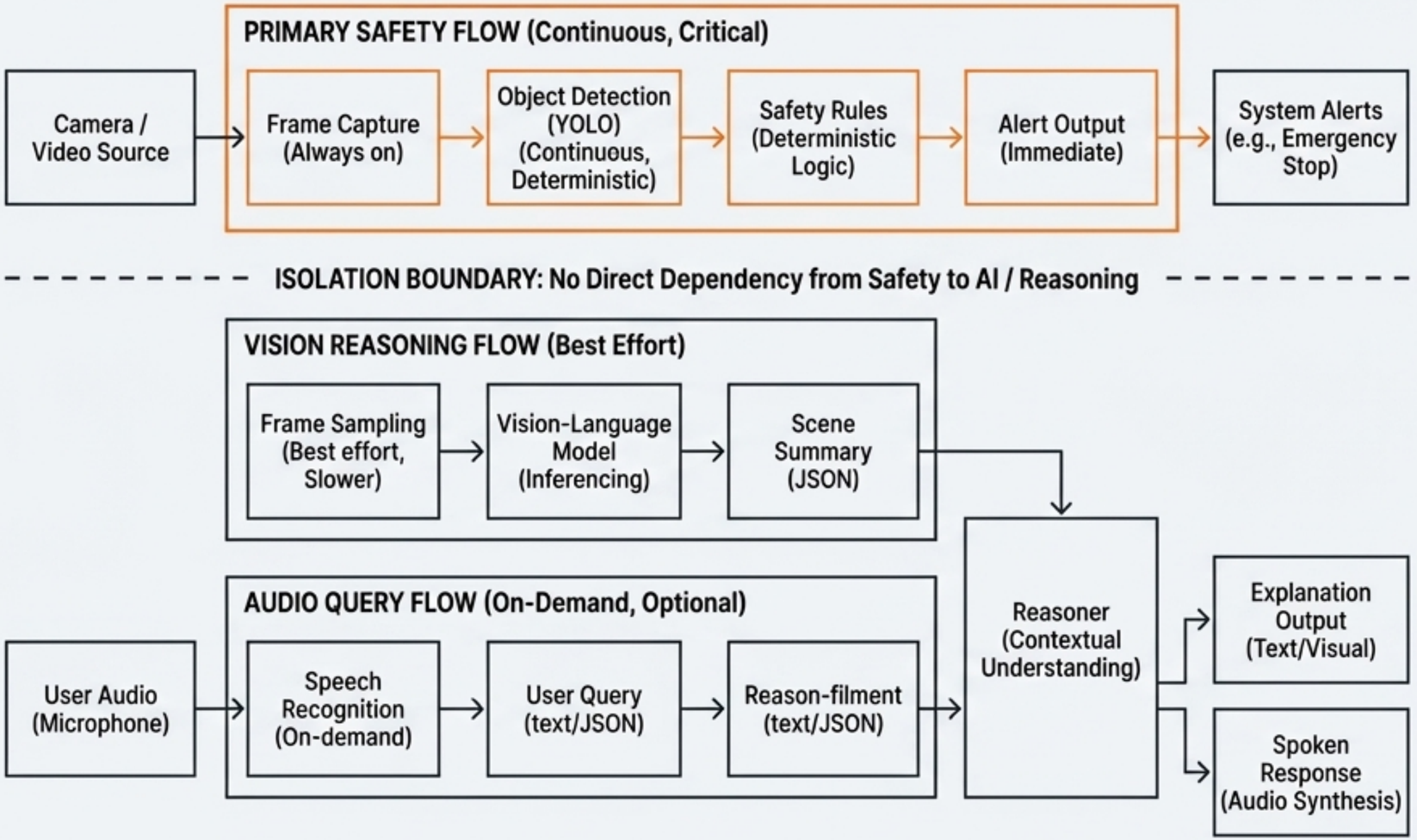
A Safety-Isolated Architecture to Balance Real-Time Perception and Semantic Reasoning

- **Perception Layer**
  Handles primary visual processing using YOLO for high-speed, deterministic object detection.
- **Reasoning Layer**
  Employs a VLM (Qwen3-VLM) for 'best-effort' contextual understanding from sampled frames.
- **Safety Layer (Isolated)**
  A critical component. Rule-based logic acts on YOLO's output to generate priority alerts, completely independent of the slower, non-deterministic VLM.
- **Orchestration & Interaction Layers**
  Fuse inputs, manage context, and handle user I/O, always prioritizing safety alerts.



Design Rationale: This layered, isolated approach ensures that critical safety functions are never blocked or delayed by computationally intensive AI reasoning tasks.

Three Parallel Processing Flows Ensure Deterministic Safety and Rich Context



Flow 1: Primary Safety Flow (Continuous, Critical)

- **Path:** Frame Capture → Object Detection (YOLO) → Safety Rules → Immediate Alert Output.
- **Characteristics:** Always on, deterministic, low-latency. Operates at full frame rate.
- **Key Principle:** Strict isolation boundary ensures no dependency on AI reasoning for safety.

Flow 2: Vision Reasoning Flow (Best Effort)

- **Path:** Frame Sampling → VLM → Scene Summary → Contextual Reasoner → Explanation Output.
- **Characteristics:** Slower, operates on sampled frames to manage computational load. Provides deep semantic understanding.

Flow 3: Audio Query Flow (On-Demand, Optional)

- **Path:** User Audio → Speech Recognition → User Query → Reasoner → Spoken Response.
- **Characteristics:** User-initiated for interactive Q&A.

The Technology Stack is Selected for Real-Time Performance on Consumer-Grade Hardware

Core Models

Object Detection

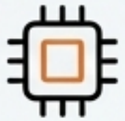
YOLOv8n (3.2M parameters)

Rationale: Chosen for its high speed (≥ 30 FPS target) and small footprint, making it ideal for the safety-critical loop.

Vision-Language Model

Qwen3 Vision–Language Model (Qwen3-VL)

Rationale: Selected for its strong scene understanding and reasoning capabilities, running asynchronously in the “best-effort” flow.



System & Hardware

OS: Windows 10/11
CPU: Quad-core (Intel i5 / Ryzen 5)
GPU: NVIDIA (≥ 6GB VRAM rec.)
RAM: ≥ 16 GB

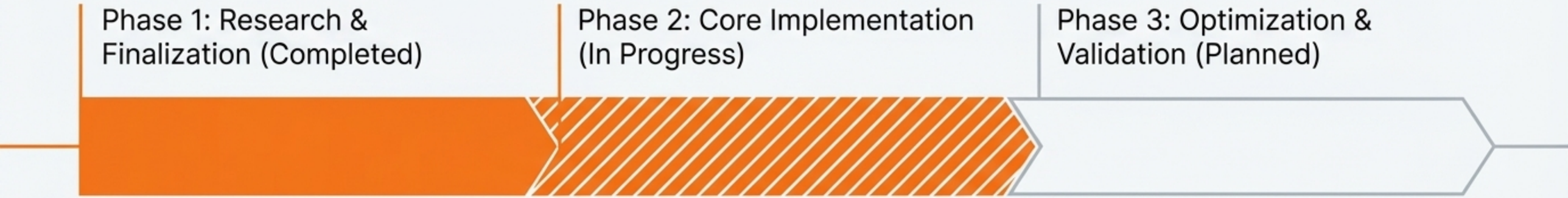
Audio Pipeline

STT: SpeechRecognition library
TTS: pyttsx3 (offline synthesis)
Latency Target: Sub-second response for alerts

Deployment

Execution: Fully local, on-device
Optimization: ONNX export, quantization planned

Project Status: Core System Architecture and All Key Modules are Implemented



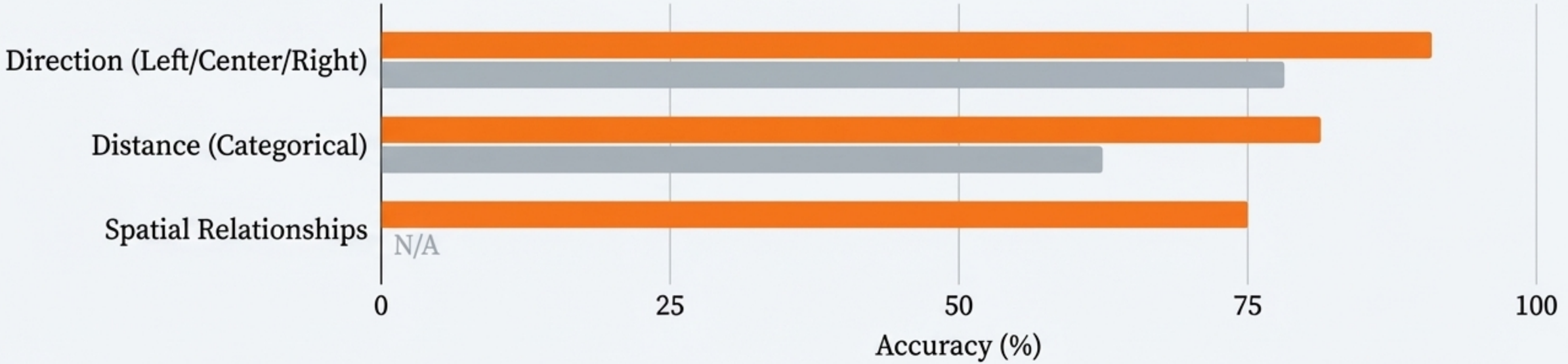
- Literature Review
- System Architecture & Tech Stack Finalization
- Dataset Preparation

- Integration of vision (YOLO), VLM (Qwen), OCR, and speech models.
- Establishment of the main processing engine and interaction controller.

- Real-Time Execution & Optimization (ONNX/TFLite conversion)
- Testing & User Feedback Loop
- Final Refinements & Documentation

Key Finding #1: VLM Demonstrates Superior Spatial Accuracy Over Heuristic Methods

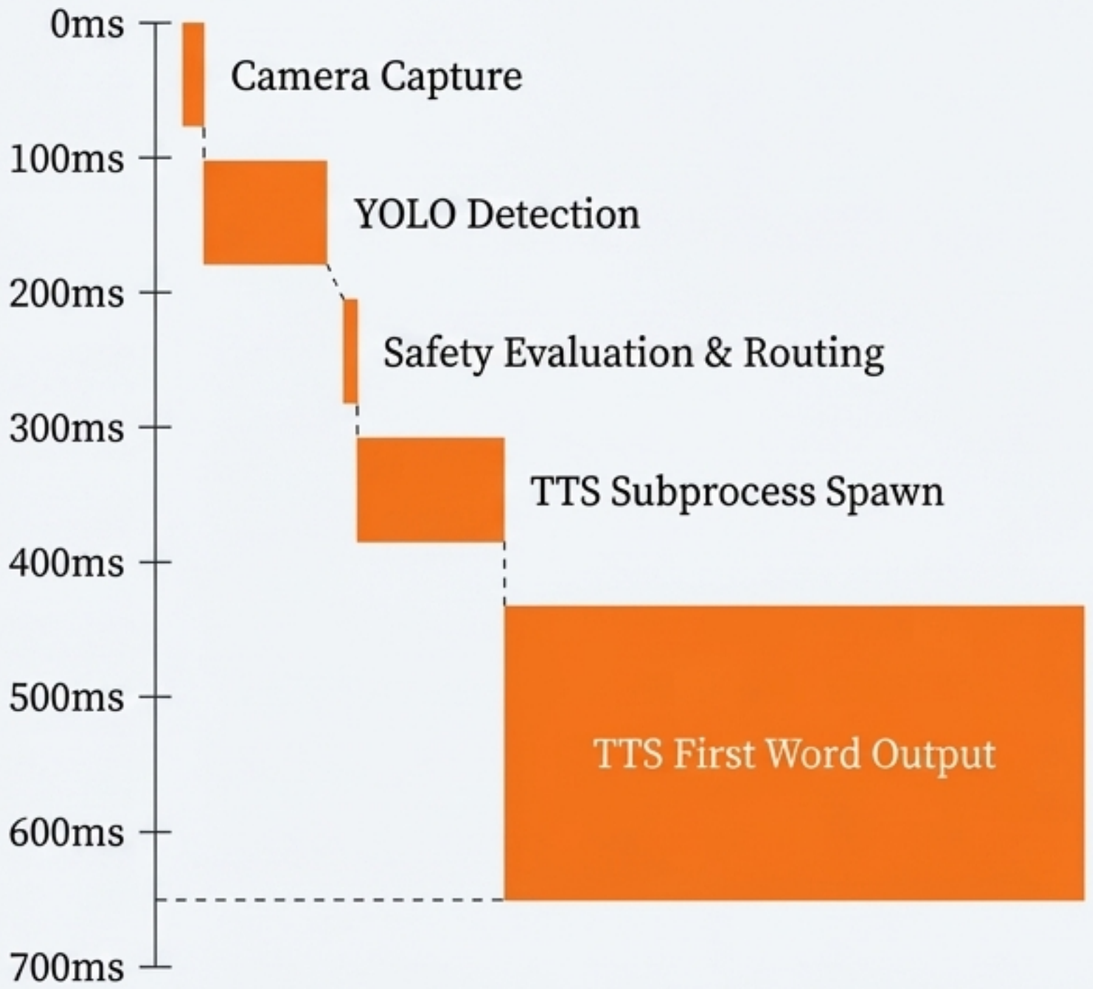
Context: We empirically tested the accuracy of spatial reasoning using two methods: a heuristic based on YOLO's bounding box position ("YOLO Method") and direct querying of the VLM ("VLM Method").



Metric	YOLO Method (Heuristic)	VLM Method (Reasoning)	Winner
Direction (Left/Center/Right)	70-80%	**85-95%**	**VLM** ✅
Distance (Categorical)	60-70%	**75-85%**	**VLM** ✅
Spatial Relationships	N/A	**70-80%**	**VLM** ✅

Implication: This validates our hybrid approach. We use YOLO for its speed in the safety loop, but rely on the slower, more accurate VLM for answering detailed spatial user queries.

Key Finding #2: The Critical Safety Path Delivers Alerts from Detection-to-Speech in Under 650ms



****End-to-End Latency Breakdown (Critical Path)**

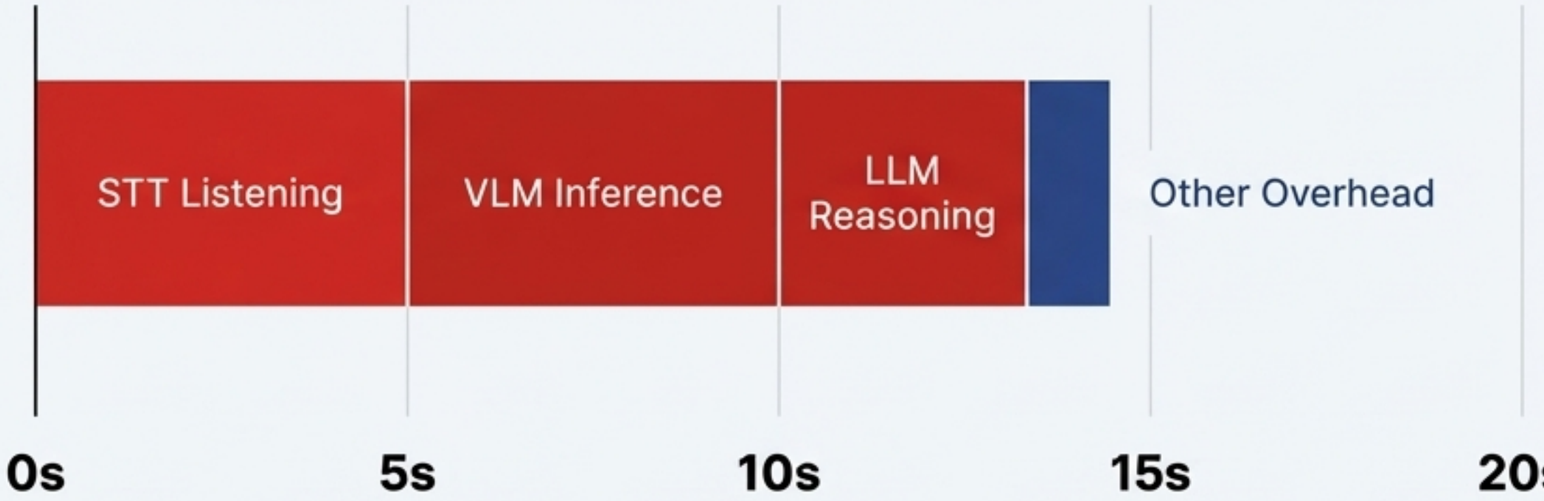
Stage	Latency	Cumulative
Camera Capture	1-5ms	1-5ms
YOLO Detection	30-50ms	31-55ms
Safety Evaluation & Routing	<2ms	32-57ms
TTS Subprocess Spawn	50-100ms	82-157ms
TTS First Word Output	200-500ms	**282-657ms**

Conclusion: Total latency is **280-650ms.** The system successfully meets the real-time requirement for its most critical function, proving the effectiveness of the safety-isolated architecture.

Technical Editorial

Identified Challenge: The Full VLM+LLM Pipeline Introduces Significant User Query Latency

Context: While the safety path is fast, queries requiring fresh visual analysis and language model reasoning are currently very slow, impacting the user experience.



Latency Breakdown (Slowest Path)):

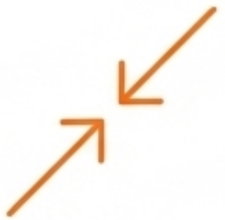
- **STT Listening:** 2.0 – 5.0 seconds
- VLM Inference: 2.0 – 5.0 seconds
- LLM Reasoning: 1.0 – 3.0 seconds
- Other Overhead (TTS, etc.): ~0.5 – 1.2 seconds

Total Latency: 5.5 to 19.2 seconds.

Implication: This performance is unacceptable for a fluid interactive experience. This bottleneck is a primary focus for our future optimization work.

Rigorous Self-Critique Has Identified Key Architectural Collisions for Refinement

Introduction: Beyond performance metrics, we've identified several logic and state management conflicts that need to be addressed to improve system robustness.



1. Query State Synchronization Issue

The user query state is tracked in two separate locations, creating a race condition that can lead to duplicate processing or missed queries.

2. Dual Redundancy Checking

Two independent redundancy filters (one for TTS output, one for VLM history) exist with different thresholds, causing inconsistent message suppression.

3. STT Blocking

Despite threading, the current Speech-to-Text implementation blocks the main loop for 2-5 seconds during listening, freezing real-time detection.

4. Spatial TTS Interrupts

Automatic announcements for nearby objects (a WARNING) can interrupt a user-requested scene description or query answer (a RESPONSE), violating conversational priority.

A Strategic Roadmap to Address Key Gaps and Enhance System Intelligence

Phase 1: Optimization & Architectural Fixes

Goal: Address Latency and Robustness.

Actions:

- Implement **Fast Query Routing** to answer simple spatial questions in <500ms, bypassing the slow VLM path. (Addresses Query Latency)
- Consolidate redundancy checking and centralize query state. (Fixes Architectural Collisions)
- Implement non-blocking Speech-to-Text.

→

Phase 2: Enhancing Intelligence & Accuracy

Goal: Improve Contextual Understanding.

Actions:

- Integrate a **Monocular Depth Estimation Model** (e.g., MiDaS) for metric distance.
- Implement a **Temporal Context Buffer** to enable collision prediction.

→

Phase 3: User Studies & Validation

Goal: Validate Usability and Effectiveness.

Actions:

- Conduct functional and user testing with visually impaired participants.
- Gather qualitative feedback and iterate on interaction design.

WalkSense Delivers a Novel and Reusable Architectural Pattern for Assistive Systems

Summary of Key Contributions

1. A Validated Hybrid Architecture

Successfully demonstrates how a real-time safety loop (**30-50ms**) can coexist with a deep semantic reasoning loop (**2-5s**) without compromising critical alert speed.

2. An Adaptive Multi-Tier Query Design

A novel routing pattern that can answer **~70%** of user queries in under **<500ms** by using the fastest available data source, reserving the slow VLM/LLM path for only the most complex questions.

3. Empirical Finding on VLM Spatial Accuracy

Provides clear evidence that modern VLMs are superior to heuristic-based methods for spatial reasoning tasks, justifying their inclusion despite higher latency.

Broader Impact: This work provides a robust, safety-first blueprint for the next generation of AI-powered assistive technologies.