Babeș-Bolyai University, Cluj-Napoca

Faculty of Mathematics and Computer Science

Real-Time Scene-Aware Assistant

for Blind People using LiquidAI

MIRPR Project Report

Team Members:

Moglan Călin

Pușcaș Raul

2025–2026

Abstract

This project proposes the development of a **Real-Time Scene-Aware Assistant** for blind or visually impaired individuals using **LiquidAI LFM2-VL** — a compact Vision-Language Model optimized for edge deployment.

The core contributions of our work include:

- **Main Idea:** A system that performs continuous real-time analysis of the environment, generating context-aware natural language descriptions.
- **Methods:** Utilization of efficient AI models (Moondream, LiquidAI LFM2) deployable on mobile or wearable devices, emphasizing autonomy and safety.
- **Data:** Experiments focus on recognition accuracy using datasets like Recaptcha-v2 and VQA-v2.
- **Results:** We demonstrate that small-scale VLMs can achieve $\sim 90\%$ accuracy in safety-critical tasks like crosswalk detection while running on low-power hardware.

Contents

Chapter 1

Introduction

1.1 What? Why? How?

This project addresses the challenge of providing intelligent visual assistance to visually impaired individuals using the **LiquidAI LFM** architecture.

What is the scientific problem? Blind or visually impaired individuals lack real-time access to visual information essential for safe navigation and interaction. Tasks such as identifying obstacles, recognizing people, or understanding environmental context traditionally require human assistance or bulky equipment.

Why is it important? Enhancing independence and safety for visually impaired users directly impacts accessibility and social inclusion. Furthermore, demonstrating that advanced multimodal reasoning can operate within limited hardware constraints (edge devices) is a significant step forward for efficient AI.

What is your basic approach? We design an application that continuously processes camera input, interprets scenes through a compact Vision-Language Model, and provides real-time natural language descriptions or spoken responses.

1.2 Paper Structure and Original Contribution(s)

The research presented in this paper advances the practical application of Vision-Language Models on edge devices.

The **main contribution** of this report is to present an intelligent algorithm for real-time scene understanding that balances accuracy with computational efficiency, achieving $\sim 90\%$ accuracy on street object detection tasks.

The **second contribution** consists of building an intuitive, voice-activated software application that allows users to query their surroundings naturally (e.g., "Is it safe to cross?").

The **third contribution** consists of a comparative analysis of different VLM architectures (Moondream vs. LiquidAI) for this specific domain.

The work is structured as follows: Chapter 2 defines the scientific problem. Chapter 3 reviews related work. Chapter 4 details our experimental approach and results. Chapter 5 describes the application study case, and we conclude in Chapter 6.

Chapter 2

Scientific Problem

2.1 Problem Definition

The core task is **Open-Vocabulary Street Object Detection and Safety Classification**. Existing systems often rely on cloud processing, leading to latency and privacy concerns, or simple bounding-box detection which lacks semantic context. An intelligent VLM algorithm is required to not just "detect" objects but "understand" the safety implications of a scene.

Formal Definition:
Let $I \in \mathbb{R}^{H \times W \times 3}$ be an input image frame. Let V be a pre-defined vocabulary of relevant traffic entities, specifically $V = \{\textit{zebra crossing}, \textit{traffic light (red)}, \textit{traffic light (green)}, \textit{none}\}$.
The objective is to learn a mapping function $f_\theta : I \to (C, S)$ parameterized by weights θ, where:

- $C \in V$ is the **Object Classification** label.
- $S \in \{\textit{Safe}, \textit{Unsafe}, \textit{Caution}\}$ is the **Safety Status**.

Input: Continuous video feed from a wearable camera and optional voice queries.
Output: Natural language descriptions and safety alerts.

Performance Criteria

We evaluate the model using standard classification metrics:

- **Accuracy (Global):** The ratio of correctly predicted observations to total observations.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

- **Precision:** The ratio of correctly predicted positive observations to the total predicted

positive observations (crucial for avoiding false alarms).

$$\text{Precision} = \frac{TP}{TP + FP}$$

- **Recall:** The ratio of correctly predicted positive observations to all observations in the actual class (crucial for safety).

$$\text{Recall} = \frac{TP}{TP + FN}$$

Chapter 3

State of the Art / Related Work

The following research highlights the advancements that inspired this project, comparing traditional methods with modern VLM approaches:

1. **YOLO (You Only Look Once) - Real-Time Object Detection**
   Goal: Fast, single-stage object detection for bounding box regression.
   Data: Trained on the **COCO Dataset** (Common Objects in Context).
   Performance: YOLOv8 achieves $\sim$ 53.9% mAP (Mean Average Precision) on COCO val2017.
   Relevance: While highly efficient for localization, it lacks the semantic understanding of scene safety required for our specific "blind assistant" task.
   Reference: Redmon et al. (CVPR 2016), https://github.com/ultralytics/ultralytics

2. **Vision Transformers (ViT)**
   Goal: Applying the Transformer architecture (self-attention) directly to image sequences of patches.
   Data: Pre-trained on **ImageNet-21k** or JFT-300M.
   Performance: Outperforms ResNet backbones on large-scale classification benchmarks ($\sim$ 88.55% top-1 accuracy on ImageNet).
   Relevance: Serves as the foundational visual encoder for modern VLMs like the one used in this project.
   Reference: Dosovitskiy et al. (ICLR 2021).

3. **Moondream (TinyVLM)**
   Goal: Creating a compact Vision-Language Model ($\sim$ 1.6B parameters) optimized for edge devices (mobile/IoT).
   Data: Trained on a mix of LLaVA, TextVQA, and synthetic data.
   Performance: Competitive with larger models (like LLaVA-7B) on VQAv2 benchmarks while being 4$\times$ faster.
   Relevance: Our initial experiments used this architecture; while promising for general Q&A, it showed instability in safety-critical classification without extensive fine-tuning.
   Reference: https://github.com/vikhyat/moondream

4. **NavVLM: Vision-Language Models for Navigation**

   Goal: Integrating VLM reasoning into robot navigation to understand spatial instructions.

   Algorithm: Uses a VLM to parse natural language commands into goal waypoints.

   Relevance: Proves the utility of language models in spatial awareness, validating our approach of using VLMs for pedestrian safety.

   Reference: Dorbala et al. (2023).

5. **Liquid Foundation Models (LFM)**

   Goal: A novel architecture based on dynamic systems and linear attention, designed to be more efficient than standard Transformers for sequence processing.

   Algorithm: Unlike static Transformers, LFMs (like LFM-1.6B) utilize adaptive dynamical systems that process inputs with significantly reduced memory footprint during inference.

   Performance: Offers state-of-the-art performance for small-scale parameters, rivaling larger Transformer models while maintaining high throughput on edge hardware.

   Relevance: This architecture serves as the backbone for our final solution, selected for its superior stability and reasoning capabilities compared to other compact VLMs.

   Reference: LiquidAI (2024), https://www.liquid.ai/

Chapter 4

Investigated Approach and Experimental Analysis

Our research followed an iterative process involving two distinct architectures: **MoondreamAI** and **LiquidAI LFM2**.

4.1 Phase 1: Experiments with MoondreamAI (Failed)

Initial attempts focused on the Moondream architecture using a hybrid training strategy with LoRA ($r = 16$) on a mixed dataset (771 local images + 500 VQA-v2 images).

Hyperparameters: LR: $1e-5$ (Local) / $3e-5$ (Hybrid), Batch Size: 1.

Table 4.1: Moondream Training Metrics

Metric	**Local-Only Run**	**Hybrid Run**
Initial Loss	3.90	6.45
Final Loss	**1.65**	3.40
Convergence	Linear (Overfitting Risk)	Unstable $\rightarrow$ Stable
Exact Match Accuracy	N/A	$\sim$ 80% (Internal Set)

Conclusion: While the hybrid model achieved stability ($\sim$ 65.8% accuracy on VQA-v2 subset), it struggled with consistent real-time inference for safety prompts, often hallucinating details.

4.2 Phase 2: Experiments with LiquidAI LFM2 (Current Success)

We transitioned to the LiquidAI LFM2-1.6B model, applying lessons learned from Phase 1.

4.2.1 Experiment 2.1: The Baseline Failure

Configuration: Learning Rate $1.5e-4$, conversational system prompt.
Outcome: Catastrophic forgetting. The model generated incoherent text loops (e.g., "python of python") and accuracy dropped to $\sim 7\%$. This confirmed that high learning rates destroy pre-trained weights in small VLMs.

4.2.2 Experiment 2.2: Stabilization (The "Safe Mode")

Adjustments:

- **Learning Rate:** Reduced to $4.0e-5$.
- **Prompt Engineering:** Enforced strict output constraints ("Output ONLY: zebra, none").
- **Repetition Penalty:** Set to 1.2.

Outcome: Loss decreased steadily from 3.5 to 1.05. Token accuracy during training reached 96%.

4.3 Final Results Analysis

The optimized LiquidAI model was evaluated on a held-out validation set of 500 images.

Table 4.2: Performance Metrics on Unseen Validation Data

Metric	Value
Global Accuracy	**89.8%**
Precision (Class: Zebra)	85.9%
Recall (Class: Zebra)	65.0%
F1-Score	0.74

Discussion: The model achieves a high Global Accuracy ($\sim 90\%$). The High Precision (86%) indicates few false alarms. The Moderate Recall (65%) suggests the model is conservative, preferring to predict "none" (safe) when uncertain, which is a specific safety behavior learned during fine-tuning.

Chapter 5

Application (Study Case)

5.1 App's Description and Main Functionalities

The application runs on mobile or wearable devices, providing real-time scene interpretation and natural language feedback.

Main Functionalities:

- Visual context description and event detection.
- Voice-activated question answering (e.g., "Describe my surroundings").
- Real-time feedback via text-to-speech.

5.2 Numerical Validation

The numerical validation of our approach is detailed in Chapter 4. We used a rigorous methodology involving training, validation, and testing splits to ensure the results are statistically significant and not a result of overfitting. The final accuracy of $\sim 90\%$ on unseen data validates the proposed TinyVLM approach for this specific domain.

Chapter 6

Conclusion and Future Work

This project demonstrates that small-scale Vision-Language Models like **LiquidAI** can deliver real-time, socially impactful AI solutions on low-power devices. The transition from a non-functional model to a 90% accurate system highlights the importance of hyperparameter tuning and data curation.

Future Work:

- Integrating multimodal sensors (audio, GPS) to improve situational awareness.
- Enhancing dialogue memory for more natural, multi-turn interaction.
- Improving Recall by augmenting the dataset with low-light and occluded examples.