Safety-Prioritised Multimodal Navigation for Edge Smart Glasses with Constrained On-Device Language Models

Şefki Kolozali, Alireza Kalantarhormozi, Abdullah Ahmad, Ali Radwan, Ahmed Jafer,
Soran Ghaderi, Mustafa Mustafa, Vishwanathan Mohan, Klaus McDonald-Maier
University of Essex, United Kingdom
Emails: sefki.kolozali@essex.ac.uk, vishwanathan.mohan@essex.ac.uk, kdm@essex.ac.uk

***Abstract*—AI-enabled assistive navigation for visually impaired individuals must satisfy strict safety, latency, and computational constraints on wearable edge platforms. This paper presents a safety-prioritised multimodal navigation framework implemented on a fully custom-designed and 3D-printed smart glasses platform integrated with an NVIDIA Jetson Orin Nano edge computing module. The system combines lightweight visual object detection, binaural audio localisation, and QR-based contextual cues with constrained small-scale large language model (LLM) reasoning. A deterministic finite state machine (FSM) enforces a hazard-first decision hierarchy, ensuring that safety-critical detections override contextual navigational suggestions while maintaining real-time operation entirely on-device.**

Experimental results demonstrate 99% audio event recognition accuracy and up to 86.5% precision 84.8% recall for object detection under edge constraints. The 0.5B parameter LLM achieves approximately 2.1 seconds inference latency on the Jetson Orin Nano. The integrated system was validated through a structured 9-minute 45-second indoor navigation trial involving dynamic and static obstacles, during which no unsafe actions occurred. These findings show that compact multimodal models, when embedded within a deterministic safety hierarchy and deployed on dedicated edge hardware, enable reliable real-time assistive navigation without reliance on cloud infrastructure.

I. Introduction

Visual perception is fundamental to spatial awareness and navigation. However, at least 2.2 billion people worldwide experience some form of visual impairment, limiting safe mobility and independent interaction with their surroundings [1]. Reduced mobility is associated with increased safety risks, social isolation, restricted employment opportunities, and diminished quality of life [2]. These challenges motivate the development of intelligent assistive technologies capable of reliable real-time environmental understanding under practical deployment constraints. Traditional tools such as smart canes [3] and GPS-based systems [4] provide partial support but lack continuous perception and contextual decision-making. Recent advances in deep learning enable real-time object detection on embedded hardware using lightweight models such as YOLO [5]; however, vision-only systems remain sensitive to illumination changes, depth ambiguity, and dynamic environments [6], [7]. Binaural audio processing provides complementary spatial awareness using interaural cues (ILD, ITD, IPD) [8], improving localisation when informative acoustic signals are present, but it cannot independently provide reliable object-level understanding. Assistive navigation therefore requires structured interpretation of multimodal inputs. Multimodal large language models (MLLMs), including Flamingo [9], PaLM-E [10], and BLIP-2 [11], demonstrate strong cross-modal reasoning capabilities. Compact Transformer-based models such as QWEN [12] enable constrained reasoning on edge devices, but generative models are inherently probabilistic and may produce outputs that conflict with real-world hazards, necessitating explicit control in safety-critical applications.

To address these challenges, we propose a *safety-prioritised multimodal assistive navigation framework* that integrates visual detection, binaural audio localisation, QR-based cues, and constrained language reasoning within a deterministic control architecture. The system operates entirely on-device using an NVIDIA Jetson Orin Nano, enabling low-latency operation, privacy preservation, and independence from network connectivity. A finite state machine enforces hazard-first behaviour by overriding navigation outputs when risks are detected. Unlike prior work, the proposed approach unifies multimodal perception with safety-aware control and constrained on-device reasoning. The primary contribution lies not in individual model performance, but in the system-level integration of these components under edge constraints. The key contributions of this work are:

- **Safety-prioritised multimodal fusion:** Decision-level integration of visual, audio, and QR inputs within a deterministic finite state machine (FSM) enforcing hazard-first override.
- **Constrained on-device reasoning:** Deployment and evaluation of compact LLMs (QWEN 0.5B and 1.8B), showing smaller models can outperform larger ones under edge constraints.
- **Separation of reasoning and control:** LLM outputs are validated by a deterministic controller prior to execution, preventing unsafe actions.
- **System-level validation:** A structured indoor navigation trial demonstrating reliable multimodal perception and consistent hazard-first behaviour.

The remainder of the paper is organised as follows. Sec-

tion II reviews existing work. Section III presents the proposed architecture. Section IV details the evaluation and results. Section V discusses implications and limitations. Section VI concludes the paper.

II. Related Work

Vision-Based Assistive Navigation

Research on smart glasses for assistive navigation has expanded significantly in recent years. Early work explored augmented reality wearables across healthcare and industrial domains, highlighting both their potential and integration challenges [13]. For visually impaired assistance, many systems rely primarily on visual perception. For example, LytNetV2 [14] employed a MobileNetV3-based architecture to detect traffic lights, achieving 96% accuracy under daytime conditions. Similar smart glasses and intelligent cane systems have also been proposed [15], but these approaches typically depend on vision alone and lack multimodal awareness.

Object detection-based systems frequently utilise YOLO architectures [5], [7]. While effective in controlled settings, they often face challenges in real-time deployment due to limited depth awareness, sensitivity to illumination, and reduced robustness in cluttered environments [6]. Joshi et al. [7] combined YOLOv3 with ultrasonic sensing to improve reliability; however, deployment on resource-constrained hardware raises concerns regarding computational load and power efficiency.

To improve spatial understanding, wearable systems integrating RGB-D cameras, GPS, and IMU data have been proposed [16]–[18]. Localisation and mapping techniques such as ORB-SLAM and ORB-SLAM3 provide strong real-time pose estimation and loop closure [19]. Assistive navigation systems have combined SLAM with semantic or instance segmentation to enhance indoor robustness [20], [21], while Transformer-based approaches further improve segmentation performance [22], [23]. However, their computational demands and reliance on external resources remain limitations for wearable deployment.

In contrast, QR-code-based indoor navigation systems provide deterministic localisation anchors. Idrees et al. [24] proposed a QR-based routing framework using markers for stepwise guidance, while Fraga et al. [25] extended this concept with computer vision for path optimisation. Although effective for structured navigation, these approaches generally focus on path following rather than multimodal perception and safety-aware decision-making.

Binaural Audio and Acoustic Localisation

Beyond vision, auditory perception plays a critical role in spatial orientation for visually impaired users. Sound source localisation (SSL) techniques exploit interaural differences such as ILD, ITD, and IPD to estimate direction and distance [8]. Binaural localisation methods have been applied in robotics and assistive systems to enhance environmental awareness, particularly in scenarios where visual cues are degraded or occluded. Map-less indoor navigation using audio cues was explored in PathFinder [26], demonstrating infrastructure-free operation but lacking integrated object detection and multimodal decision-making. Despite progress in SSL and binaural modelling, integration with visual perception and higher-level reasoning in wearable systems remains limited.

Multimodal Reasoning and Language Models

Recent multimodal large language models (MLLMs) such as Flamingo [9], PaLM-E [10], and BLIP-2 [11] have demonstrated strong cross-modal scene understanding and contextual reasoning capabilities, enabling joint visual grounding and language-based inference across complex environments. Compact Transformer architectures such as QWEN [12] incorporate efficiency-oriented mechanisms including SwiGLU activations [27], group query attention, and sliding-window attention, enabling deployment on embedded platforms such as NVIDIA Jetson devices. However, the integration of generative reasoning into safety-critical wearable navigation remains largely unexplored, particularly under strict on-device constraints.

Positioning of This Work

Prior work typically addresses perception accuracy, localisation, or scalability in isolation. SLAM-based systems focus on continuous mapping, while QR-based approaches emphasise structured navigation. Few frameworks explicitly integrate multimodal perception with constrained language reasoning under safety-critical constraints. In particular, deterministic control mechanisms that enforce hazard-prioritised overrides over generative outputs remain underexplored. This work addresses this gap through a safety-prioritised multimodal framework combining lightweight perception, QR-based context, and constrained on-device reasoning under explicit safety validation.

III. Methodology

A. System Overview

The proposed system is implemented on a custom 3D-printed smart glasses platform integrating a ZED Mini stereo depth camera, Soundman OKM II binaural microphones, and an NVIDIA Jetson Orin Nano edge computing module. Figure 1 illustrates the overall architecture, in which visual, audio, and QR-based contextual inputs are processed on-device before being passed to a deterministic decision layer. The stereo camera is centrally mounted to align with the user's forward visual axis and provide depth-aware RGB perception, while the binaural microphones are symmetrically positioned to preserve interaural cues required for spatial audio localisation. The hardware configuration is presented in Figure 2, emphasising stable sensor geometry and reproducible multimodal perception. The Jetson Orin Nano executes perception, fusion, reasoning, and decision modules locally, enabling privacy-preserving assistive navigation without cloud dependency.

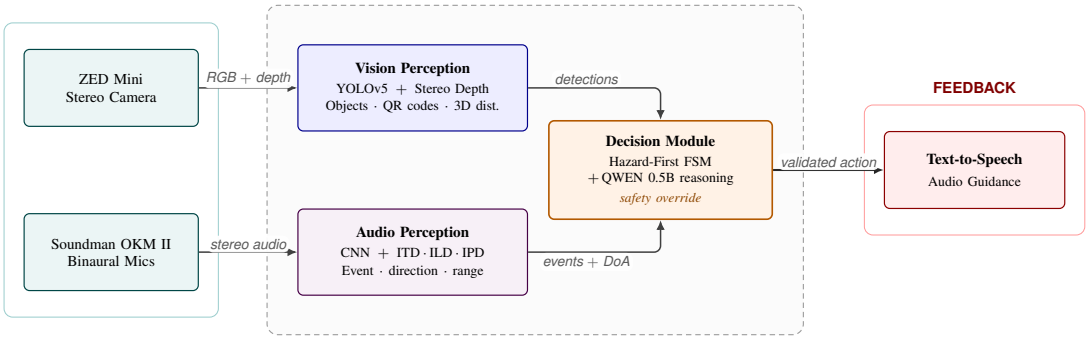


Fig. 1. End-to-end architecture of the proposed safety-prioritised multimodal smart-glasses navigation system. Stereo vision and binaural audio streams are processed in parallel by lightweight perception modules on the Jetson Orin Nano. A deterministic finite-state machine fuses detections with QWEN-based contextual reasoning under a hazard-first priority, and validated actions are delivered to the user via on-device text-to-speech.



(a) Smart glasses prototype. (b) Jetson Orin Nano enclosure.

Fig. 2. Hardware platform for multimodal assistive navigation: (a) custom-designed and 3D-printed smart glasses integrating the ZED Mini stereo depth camera and Soundman OKM II binaural microphones; (b) custom-designed and 3D-printed enclosure for the NVIDIA Jetson Orin Nano.

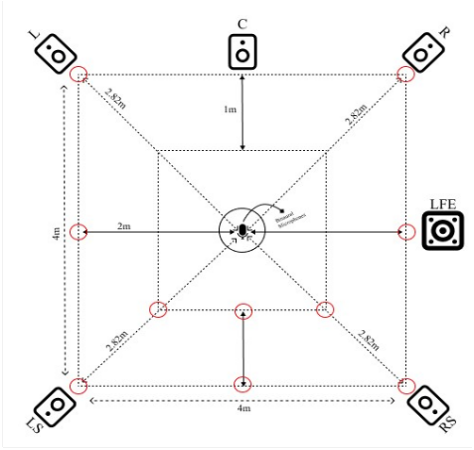


Fig. 3. Spatial configuration of the 5.1 surround recording setup in the 4 m × 4 m robotics arena. Red markers indicate the ten spatial sound emission locations used to generate direction–distance combinations.

B. Binaural Audio Dataset and Recording Protocol

Audio samples were derived from the publicly available *Making Sense of Sounds (MSoS)* dataset. Eighteen environmental sound classes relevant to indoor assistive navigation were selected, including hazardous sounds (e.g., glass breaking, door slams, object drops), human activity sounds (e.g., footsteps, speech, clapping), and ambient mechanical sounds (e.g., alarms, sirens, elevator sounds, trolley movement). To emulate realistic spatial acoustic conditions, mono audio samples were spatialised using Cubase 12 Professional and a 5.1 surround configuration via the VST Multipanner tool. The spatialised audio was reproduced inside a controlled 4 m × 4 m robotics arena using six loudspeakers positioned at fixed azimuth angles following standard surround nomenclature: FL (Front Left), FR (Front Right), C (Centre), LS (Left Surround), RS (Right Surround), and LFE (Low-Frequency Effects channel), as illustrated in Fig. 3.

The binaural microphones mounted on the smart glasses were positioned centrally within the arena. Each of the 18 sound classes was replayed from ten predefined spatial emission locations, generating diverse direction–distance combinations. The sounds were physically propagated through the acoustic environment and re-recorded using the binaural microphones, ensuring that interaural cues resulted from real acoustic propagation rather than synthetic stereo manipulation. In total, 3380 binaural recordings were generated.

C. Audio Feature Extraction and Task Formulation

Audio signals were sampled at 48 kHz and processed in fixed-length segments. Mel-spectrograms were computed independently for left and right channels and used as the primary spectral representation. Interaural Time Difference (ITD) was estimated via cross-correlation between channels to capture arrival-time disparities. Interaural Level Difference (ILD) was computed from log-energy differences between channels to represent intensity variation, and Interaural Phase Difference (IPD) was derived from STFT-based phase differences. These cues encode horizontal localisation and distance-related spatial characteristics. Three audio prediction tasks were formulated: (i) 18-class environmental sound recognition, (ii) 3-class direction estimation (Left, Right, Back), and (iii) binary distance estimation (Near, Far).

All tasks were implemented using task-specific convolutional neural network (CNN) architectures. The *event recognition model* processes concatenated left–right Mel-spectrograms using a convolutional backbone consisting of a 3×3 convolutional layer with 32 filters (ReLU), followed by

2×2 max pooling and dropout (0.5), and two additional 3×3 convolutional layers with 64 filters. Feature maps are passed to fully connected layers of 32 and 18 units for classification, trained using Adam optimisation and sparse categorical cross-entropy. The *direction estimation model* shares the same CNN backbone for spectral features, while interaural features (ITD, ILD, IPD) are processed through a parallel dense layer with 8 units. The two branches are concatenated and passed through a 16-unit dense layer followed by a 3-class output layer. The *distance estimation model* uses a deeper convolutional architecture with four strided convolutional layers (64 filters, 4×4 kernels, stride 2, batch normalisation, ReLU). Interaural features are processed using a multilayer perceptron and fused with CNN features prior to an Atrous Spatial Pyramid Pooling (ASPP) module with dilation rates of 6, 12, and 18. The final classifier consists of a 32-unit dense layer and a sigmoid output for Near/Far prediction. All non-output dense layers use ReLU activations.

D. Visual Object Detection

Visual perception was implemented using lightweight YOLOv5 variants trained on eight indoor navigation-relevant classes: person, stairs, door, elevator, wet-floor sign, chair, garbage container, and backpack. YOLOv5 was selected due to its proven compatibility with the Jetson platform and availability of multiple model sizes enabling systematic accuracy–efficiency trade-off analysis. For each frame, the detector outputs bounding boxes and confidence scores. Object proximity is estimated using stereo depth measurements from the ZED Mini, supported by bounding box location and scale. Objects detected within a 1 m proximity threshold are treated as potential hazards and trigger avoidance guidance through the decision-making module. Consecutive frame detections are analysed to monitor changes in object proximity, enabling response to dynamic pedestrians and moving obstacles without explicit motion modelling.

The training dataset was constructed from two sources: (i) video footage collected using the smart glasses platform along predefined campus routes, from which frames were extracted, and (ii) curated images collected from publicly available datasets, including Roboflow repositories, Mendeley datasets, and MS COCO, filtered to match the target object classes. All images were manually annotated with bounding boxes, and data augmentation (flipping, rotation, shearing) was applied using Roboflow to improve diversity and generalisation. All models were fine-tuned from pre-trained weights under consistent hyperparameters to ensure fair comparison. YOLOv5 models used an input resolution of 400×400, while SSD baselines used 512×512, with a batch size of 4 and 25 training epochs. The dataset distribution across training, validation, and test splits is shown in Table I.

TABLE I
DATASET DISTRIBUTION FOR VISUAL OBJECT DETECTION.

Class	Train	Val.	Test
Person	862	252	114
Door	1305	414	171
Elevator	1090	342	113
Stairs	1070	310	155
Garbage Container	1697	525	235
Wet Floor Sign	728	225	97
Backpack	1056	336	136
Chair	896	270	97

E. Multimodal Hazard Handling and Decision Control

Objects detected within a 1 m proximity threshold are treated as hazards and trigger immediate avoidance guidance. Hazardous audio events are prioritised similarly, with navigation instructions suppressed while a threat is active. QR-based localisation provides structured guidance once hazards are cleared. Final action selection is governed by a deterministic FSM that integrates visual detections, proximity checks, audio events, and QR metadata. Multimodal fusion is performed at the decision level, where visual (object and depth), audio (event and direction), and QR signals are independently computed and combined as discrete inputs to the FSM. These inputs are evaluated in a fixed priority order: (1) audio hazards, (2) visual proximity violations, (3) QR navigation, and (4) LLM-based guidance, ensuring that safety-critical signals override contextual navigation. Vision provides object identity and metric depth, audio captures non-visual hazards and directional cues, and QR codes provide structured localisation. The 1 m threshold defines a conservative reaction margin, while the binary Near/Far formulation reduces ambiguity in reverberant conditions. Precise proximity estimation is handled by stereo vision, with audio distance used only as supplementary context. This integration enables detection redundancy, where hazards missed by one modality can still be captured by another, improving overall system robustness.

F. LLM-Based Contextual Reasoning

Contextual reasoning is implemented using instruction-tuned QWEN 0.5B and 1.8B models based on decoder-only Transformer architectures. The 0.5B model consists of 24 Transformer layers with multi-head self-attention and feed-forward blocks, while the 1.8B variant increases depth and hidden dimensionality. To enable edge deployment, both models are quantised to 8-bit precision to reduce memory footprint. Structured multimodal state representations—including detected objects, hazard confidence, stereo depth, audio events, and QR metadata—are encoded within a fixed context window. The LLM generates candidate navigation actions in structured JSON format constrained to a predefined action vocabulary and does not directly control system outputs.

The prompt follows a two-role instruction format. The *system* role constrains outputs to a predefined action set: `["descend stairs", "ascend stairs", "stop", "continue", "turn right", "turn left"]`. The *user* role encodes the current multimodal state as a structured natural-language description combining perception outputs (e.g., object class, depth, hazard confidence, audio cues, and QR information). For example: `"Detected:`

`Wet floor at 0.6m; hazard confidence high; QR indicates forward.`" The model returns a JSON object containing a single action, ensuring consistent and constrained navigation decisions.

Hierarchical Decision-Making Framework

Final action selection is governed by a deterministic FSM acting as the primary safety controller. A binary hazard indicator $\lambda_t \in \{0, 1\}$ is computed using conservative rule-based checks: $\lambda_t = 1$ if any detected object lies within a 1 m proximity threshold or if a hazardous audio event is detected; otherwise $\lambda_t = 0$. Action selection follows:

$$A_t = \begin{cases} A_{\text{hazard}}, & \text{if } \lambda_t = 1, \\ A_{\text{navigation}}, & \text{otherwise.} \end{cases} \tag{1}$$

As shown in (1), hazard-related actions override navigation suggestions. LLM outputs are treated as candidate actions and are validated against safety constraints before being issued to the user. The FSM operates over five states—hazard audio, hazard visual, QR-based navigation, LLM-based navigation, and error handling—evaluated in strict priority order at each processing cycle, as shown in the FSM flow diagram in Fig. 4. Hazard states suppress navigation outputs, while QR and LLM states provide guidance only when no hazard is present. This hierarchical structure ensures reproducibility and enforces a clear separation between deterministic safety control and probabilistic reasoning. Generative language models may produce outputs that conflict with physical safety constraints. For example, when presented with a wet-floor hazard at 0.6 m, the QWEN 1.8B model generated "`move forward`". The FSM mitigates such failures by intercepting and rejecting unsafe actions before execution, ensuring that navigation decisions remain consistent with real-world safety requirements.

G. Edge Deployment

All perception, fusion, reasoning, and decision modules operate on the NVIDIA Jetson Orin Nano device. Visual and audio inference execute in parallel, while LLM reasoning is invoked conditionally after safety evaluation. This architecture supports wearable edge deployment under computational and privacy constraints. Real-time safety response is defined as detecting hazards and issuing corrective guidance within a timeframe that allows a pedestrian to react before contact, targeted at under 3 s for obstacle avoidance. This threshold represents a system-level design requirement rather than a measured end-to-end latency result. Safety-critical responses are handled by the audio–visual perception and FSM layers without requiring LLM inference. The measured QWEN 0.5B latency of approximately 2.1 s is therefore restricted to non-critical contextual navigation guidance. A full end-to-end latency breakdown was not instrumented and is treated as a limitation; however, the architecture ensures that LLM latency does not delay hazard override decisions.

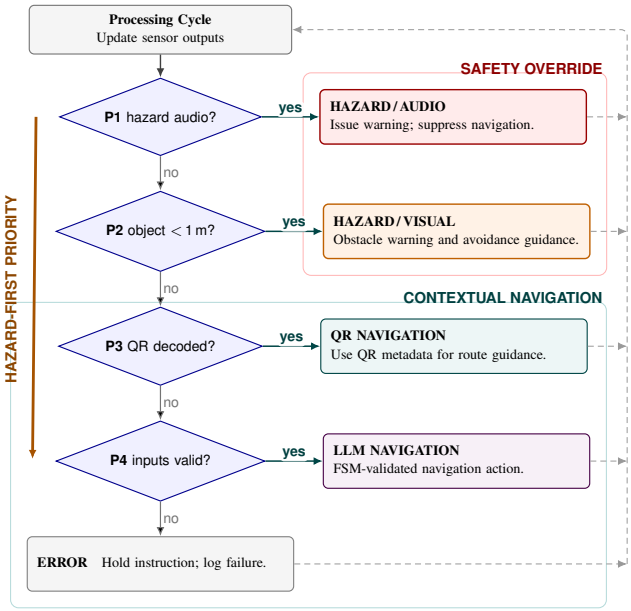


Fig. 4. Finite state machine used by the decision-making module. Sensor outputs are evaluated in strict priority order: hazardous audio, hazardous visual obstacle, QR-based navigation, LLM-based navigation, and error handling.

IV. Experimental Evaluation

A. Evaluation Protocol

The proposed system was evaluated at three levels: (i) unimodal perception performance for audio and vision modules, (ii) constrained LLM reasoning under structured prompts, and (iii) integrated system behaviour in a controlled indoor navigation scenario. Audio models were trained using an 80% training split with five-fold cross-validation, with final performance reported on a held-out 20% test set. Visual object detection models (SSD and YOLO variants) were fine-tuned from pre-trained weights under consistent hyperparameters (25 epochs with fixed architecture-specific input resolutions) to ensure fair comparison. YOLOv5 models were initialised from COCO weights, while SSD models used ImageNet-pretrained backbones. Audio models were trained from scratch on the binaural MSoS-derived dataset.

The LLM component was evaluated using 20 structured navigation prompts derived from multimodal system states, with response latency measured directly on the NVIDIA Jetson Orin Nano. Precision, Recall, and F1-score were used for audio classification, while mean Average Precision (mAP) and class-wise metrics were used for object detection. LLM responses were assessed by eight independent human raters using a 4-point relevance scale (0 = incorrect, 4 = fully appropriate), with scores averaged across raters and prompts.

Integrated system behaviour was assessed through a structured corridor navigation experiment lasting 9 minutes and 45 seconds. The evaluation involved a single participant under supervised conditions and does not constitute a full user study. The participant followed system-generated instructions without external guidance; a safety observer was present but

did not intervene during the navigation task, while structured obstacles were introduced to evaluate hazard-first override behaviour.

B. Audio Event Recognition and Localisation

The performance of the separately optimised audio models is summarised in Table II. Event recognition achieved near-perfect performance (99%), direction estimation achieved 74% test accuracy using binaural cues (ITD, ILD, IPD), while distance estimation remained comparatively challenging.

TABLE II
PERFORMANCE OF SEPARATE AUDIO MODELS.

Task	Cross-Validation	Test Accuracy
Event	99%	99%
Distance	68%	68%
Direction	73%	74%

A class-wise summary of the event and direction models is shown in Table III. For clarity, only a subset of representative, safety-relevant sound classes is presented, although the model was trained and evaluated on all 18 classes. Among these, critical sounds such as footsteps, glass breaking, and screaming achieved 100% precision and recall, reinforcing the effectiveness of the audio module as a complementary modality for hazard awareness.

TABLE III
CLASS-WISE PERFORMANCE SUMMARY OF THE AUDIO EVENT AND DIRECTION MODELS.

Model / Class	Precision	Recall	F1
Event Recognition Model			
Overall	97%	98%	98%
Footsteps	100%	100%	100%
Glass breaking	100%	100%	100%
Scream	100%	100%	100%
Siren	96%	100%	98%
Train	91%	100%	95%
Water	100%	98%	99%
Direction Model			
Left	79%	56%	66%
Right	86%	57%	68%
Back	76%	92%	83%

A more detailed class-wise analysis of the distance estimation model is provided in Table IV. The results indicate a complete failure to classify the Near class, with the model effectively collapsing to a majority-class predictor.

TABLE IV
CLASS-WISE PERFORMANCE OF THE AUDIO DISTANCE MODEL.

Class	Precision	Recall	F1
Near	0%	0%	0%
Far	68%	100%	81%
Total	34%	50%	40%

This failure reflects dataset imbalance and the weak separability of near/far acoustic cues in reverberant indoor environments. In contrast, the direction model (Left/Right/Back) achieves moderate performance, demonstrating that binaural cues provide useful spatial localisation. As a result, audio distance estimation is not used in the final decision-making process; proximity-based obstacle handling is instead derived from stereo vision. The audio module is used for event recognition and directional awareness, providing complementary detection of non-visual hazards such as footsteps and alarms. Proximity-related descriptions in system outputs are derived from stereo depth measurements, ensuring consistency between perception signals and generated guidance.

C. Visual Object Detection

The performance of visual object detection models is presented in Table V. YOLOv5 variants consistently outperform SSD baselines, with YOLOv5-Medium achieving the highest accuracy.

TABLE V
COMPARISON OF OBJECT DETECTION MODELS.

Model	Precision	Recall	F1
SSD ResNet50	79.3%	62.9%	70.15%
SSD MobileNet V2 FPNLite	81.8%	64.8%	72.32%
SSD MobileNet V1 FPN	86%	70%	77.17%
YOLOv5 Nano	83.6%	79.1%	81.28%
YOLOv5 Small	86.5%	84.8%	85.64%
YOLOv5 Medium	90.1%	85.9%	87.94%

Despite the higher accuracy of YOLOv5-Medium, YOLOv5-Small provides the most suitable balance between detection performance and computational efficiency for edge deployment. YOLOv5 was selected because of its verified compatibility with the ZED Mini stereo camera SDK and NVIDIA Jetson platform, well-established training pipeline, and availability of model sizes that support systematic accuracy–efficiency comparison. While explicit frame-rate benchmarking was not conducted, deployment observations indicated stable operation for YOLOv5-Small, whereas larger models introduced increased computational overhead with limited practical benefit.

TABLE VI
CLASS-WISE RESULTS FOR THE DEPLOYED YOLOV5-SMALL MODEL.

Class	F1	mAP50	mAP50-95
Person	47.74%	43.8%	20.7%
Door	83.73%	88.4%	62.5%
Elevator	93.44%	96.6%	70.8%
Stairs	94.35%	97.1%	69.2%
Garbage Container	93.31%	97.4%	72.6%
Wet Floor Sign	98.08%	99.4%	77.1%
Backpack	85.49%	88.8%	56.7%
Chair	84.81%	90.6%	64.1%

Table VI provides class-wise results for the deployed YOLOv5-Small model. The lower performance for the Person class indicates the need for broader training diversity, particularly for dynamic and partially occluded pedestrians.

D. LLM Evaluation

The results of the LLM evaluation are summarised in Table VII. The QWEN 0.5B model achieved a high average rating of 3.85 with tightly bounded latency between 2.122 and 2.208 seconds. In contrast, the 1.8B model achieved a lower

average rating of 1.25, with higher latency variability ranging from 3.984 to 7.677 seconds.

TABLE VII
HUMAN-RATED ACTION RELEVANCE AND LATENCY.

Model	Avg. rating	Mean time (s)	Min	Max
QWEN 0.5B	3.85	2.172	2.122	2.208
QWEN 1.8B	1.25	5.682	3.984	7.677

To further interpret these results, a correctness-based analysis derived from human ratings is presented in Table VIII. Ratings of 0–1 were treated as incorrect, while ratings of 3–4 were treated as correct. The QWEN 0.5B model produced consistently correct actions across all prompts, whereas the 1.8B model generated incorrect outputs in 65% of cases. Although this does not constitute a formal statistical evaluation, it provides a practical indication of alignment between model outputs and expected navigation behaviour. Occasional conservative outputs, such as unnecessary stopping, were observed but do not negatively affect system operation due to the hazard-first control policy. This evaluation is based on structured prompts and does not constitute a fully autonomous real-world assessment, which remains an important direction for future work.

TABLE VIII
DERIVED CORRECTNESS ANALYSIS OF LLM DECISIONS.

Model	Incorrect	Correct	Error Rate
QWEN 0.5B	0 / 20	20 / 20	0%
QWEN 1.8B	13 / 20	0 / 20	65%

Example 1: Contextual navigation. For the prompt `Detected: Stairs at 0.8m; no hazards; QR level transition`, QWEN 0.5B generated `"ascend stairs"`, whereas QWEN 1.8B generated `"turn left"`.

Example 2: Hazard scenario. For the prompt `Detected: Wet floor at 0.6m; hazard confidence high; QR indicates forward`, QWEN 0.5B generated `"stop"`, whereas QWEN 1.8B generated `"move forward"`. The FSM rejected unsafe candidate actions and enforced hazard-first behaviour.

E. Integrated System Behaviour

The integrated system was evaluated in a structured indoor navigation scenario with controlled obstacle placement, including a wet-floor warning sign, a rolling chair, a side-positioned bin, a temporarily closed door, and dynamic pedestrian proximity events. The system generated context-aware hazard alerts and consistently enforced hazard-first overrides, ensuring that safety-critical responses superseded navigation instructions. QR markers provided deterministic navigation guidance, enabling successful route completion without collision. Proximity descriptions such as “very close” are derived from stereo depth measurements rather than audio distance estimation, ensuring consistency between perception and output. The multimodal design improves robustness by allowing hazards detected by one modality to be recognised when others are unreliable. Although a full ablation study was not conducted, observations during testing suggest that removing audio cues reduces awareness of non-visual hazards, while removing QR cues reduces navigation consistency. While the evaluation demonstrates stable multimodal integration under structured conditions, generalisation to more complex environments remains to be investigated. A full end-to-end latency breakdown was not instrumented; however, safety-critical responses are handled by low-latency perception and deterministic control components, while LLM reasoning is restricted to non-critical guidance tasks.

V. DISCUSSION

The evaluation demonstrates that combining lightweight perception models with a deterministic safety hierarchy improves robustness for indoor assistive navigation. While YOLO-based detection and spectrogram-based audio models achieved strong standalone performance, system reliability is primarily governed by hazard-first override control. During the 9-minute and 45-second structured corridor trial, safety interventions consistently superseded navigation suggestions, confirming effective separation between generative reasoning and deterministic safety enforcement. Structured obstacle insertion—including a wet-floor sign, rolling chair, side-positioned bin, closed-door interaction, and pedestrian proximity—validated real-time, proximity-aware behaviour, with QR-based cues enabling navigation resumption. Multimodal fusion improved robustness under occlusion and dynamic conditions, although audio distance estimation remained sensitive to reverberation.

These results indicate that robustness depends not only on individual model accuracy but on the interaction between perception modules and control logic. The FSM enforces safety constraints by intercepting invalid or unsafe LLM outputs before execution, mitigating risks associated with probabilistic reasoning in safety-critical scenarios. Compact instruction-tuned language models revealed a trade-off between latency and contextual alignment. The 0.5B model produced faster and more consistent outputs, while the 1.8B model generated a higher proportion of incorrect actions, demonstrating that larger models do not necessarily yield better performance under edge constraints. All LLM outputs were restricted to a predefined action space and validated prior to execution.

The evaluation was conducted under structured indoor conditions with a single supervised participant, limiting generalisability to fully independent real-world deployment. Visual domain variation, such as the wet-floor sign, highlights the need for more diverse training data. The high audio event recognition performance should be interpreted in the context of the controlled recording environment, requiring further validation under more diverse and unconstrained acoustic conditions. A full end-to-end latency breakdown was not instrumented; however, real-time safety is ensured architecturally, with hazard responses handled by low-latency perception and FSM components, while LLM reasoning is restricted to non-critical guidance. LLM evaluation relies on human judgement rather

than labelled datasets, although multiple independent raters provide a practical assessment of action quality. The proposed framework distinguishes itself through explicit safety-aware integration of multimodal perception and constrained on-device reasoning. A cloud-based deployment alternative was considered but not adopted due to network-dependent latency, privacy concerns, and single-point-of-failure risks. The fully on-device architecture sacrifices access to larger models but ensures availability, deterministic latency, and user privacy, which are critical for assistive navigation.

VI. Conclusion

This paper presented a safety-prioritised multimodal navigation framework for edge-based smart glasses, integrating stereo vision, binaural audio localisation, QR-based contextual cues, and constrained on-device language reasoning within a deterministic control hierarchy. By separating generative reasoning from safety enforcement, the system achieves real-time operation while maintaining explicit hazard guarantees. Experimental results demonstrate reliable multimodal perception, stable QR-based navigation, and safe constrained language reasoning under edge constraints. The findings indicate that compact language models can provide effective contextual guidance when combined with structured prompting and deterministic validation, highlighting that system-level safety is driven by architectural design rather than model complexity.

Future work will focus on large-scale user studies with visually impaired participants and improved generalisation through more diverse visual and acoustic datasets. In addition, we will conduct a detailed end-to-end latency analysis and explore deployment on more powerful edge platforms to support advanced models while preserving the safety-first control framework.

References

[1] W. H. Organization, "World health organization, world report on vision executive summary, 2019." 20.

[2] S. T. Course, "Icts in education for people with special needs," UNESCO Institute for Information Technologies in Education (IITE), Moscow, Russia, Tech. Rep., 2006.

[3] M. Bansal, S. Malik, M. Kumar, and N. Meena, "Arduino based smart walking cane for visually impaired people," in *2020 Fourth International Conference on Inventive Systems and Control (ICISC)*, 2020, pp. 462–465.

[4] M. D. Messaoudi, B.-A. J. Menelas, and H. Mcheick, "Review of navigation assistive tools and technologies for the visually impaired," *Sensors*, vol. 22, no. 20, p. 7888, 2022.

[5] J.-Y. Lin, C.-L. Chiang, M.-J. Wu, C.-C. Yao, and M.-C. Chen, "Smart glasses application system for visually impaired people based on deep learning," in *2020 Indo – Taiwan 2nd International Conference on Computing, Analytics and Networks (Indo-Taiwan ICAN)*, 2020, pp. 202–206.

[6] A. Mustafa, A. Omer, and O. Mohammed, "Intelligent glasses for visually impaired people," in *2022 14th International Conference on Computational Intelligence and Communication Networks (CICN)*, 2022, pp. 29–33.

[7] R. C. Joshi, S. Yadav, M. K. Dutta, and C. M. Travieso-Gonzalez, "Efficient multi-object detection and smart navigation using artificial intelligence for visually impaired people," *Entropy*, vol. 22, no. 9, p. 941, 2020.

[8] P.-A. Grumiaux, S. Kitić, L. Girin, and A. Guérin, "A survey of sound source localization with deep learning methods," *The Journal of the Acoustical Society of America*, vol. 152, no. 1, pp. 107–151, 2022.

[9] J.-B. Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc, A. Mensch, K. Millicah, M. Reynolds *et al.*, "Flamingo: A visual language model for few-shot learning," in *Advances in Neural Information Processing Systems*, vol. 36, 2022, pp. 23 716–23 736.

[10] D. Driess, F. Xia, M. S. M. Sajjadi, C. Lynch, A. Chowdhery, B. Ichter, A. Wahid, J. Tompson, Q. Vuong, T. Yu *et al.*, "Palm-e: An embodied multimodal language model," in *Proceedings of the 40th International Conference on Machine Learning*, 2023, pp. 8469–8488.

[11] J. Li, D. Li, S. Savarese, and S. C. H. Hoi, "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," in *Proceedings of the 40th International Conference on Machine Learning*, 2023, pp. 19 730–19 742.

[12] J. Bai, S. Bai, Y. Chu *et al.*, "Qwen technical report," 2023. [Online]. Available: https://arxiv.org/abs/2309.16609

[13] D. Kim and Y. Choi, "Applications of smart glasses in applied sciences: A systematic review," *Applied Sciences*, vol. 11, no. 11, p. 4956, 2021.

[14] S. Yu, H. Lee, and J. Kim, "Street crossing aid using light-weight cnns for the visually impaired," in *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, 2019, pp. 2593–2601.

[15] W.-J. Chang, L.-B. Chen, C.-Y. Sie, and C.-H. Yang, "An artificial intelligence edge computing-based assistive system for visually impaired pedestrian safety at zebra crossings," *IEEE Transactions on Consumer Electronics*, vol. 67, no. 1, pp. 3–11, 2021.

[16] J. Bai, S. Lian, Z. Liu, K. Wang, and D. Liu, "Virtual-blind-road following-based wearable navigation device for blind people," *IEEE Transactions on Consumer Electronics*, vol. 64, no. 1, pp. 136–143, 2018.

[17] ——, "Smart guiding glasses for visually impaired people in indoor environment," *IEEE Transactions on Consumer Electronics*, vol. 63, no. 3, pp. 258–266, 2017.

[18] J. Bai, Z. Liu, Y. Lin, Y. Li, S. Lian, and D. Liu, "Wearable travel aid for environment perception and navigation of visually impaired people," *Electronics*, vol. 8, no. 6, p. 697, 2019.

[19] C. Campos, R. Elvira, J. J. G. Rodríguez, J. M. M. Montiel, and J. D. Tardós, "Orb-slam3: An accurate open-source library for visual, visual–inertial, and multimap slam," *IEEE Transactions on Robotics*, vol. 37, no. 6, pp. 1874–1890, 2021.

[20] J. Xu, E. Dong, Q. Ma, C. Wu, and Z. Yang, "Smartphone-based indoor visual navigation with leader-follower mode," *ACM Transactions on Sensor Networks*, vol. 17, no. 2, pp. 1–22, May 2021.

[21] Y. Song, Z. Li, G. Li, B. Wang, M. Zhu, and P. Shi, "Multi-sensory visual-auditory fusion of wearable navigation assistance for people with impaired vision," *IEEE Transactions on Automation Science and Engineering*, 2023.

[22] P.-J. Duh, Y.-C. Sung, L.-Y. F. Chiang, Y.-J. Chang, and K.-W. Chen, "V-eye: A vision-based navigation system for the visually impaired," *IEEE Transactions on Multimedia*, vol. 23, pp. 1567–1580, 2021.

[23] J. Zhang, K. Yang, A. Constantinescu, K. Peng, K. Müller, and R. Stiefelhagen, "Trans4trans: Efficient transformer for transparent object segmentation to help visually impaired people navigate in the real world," in *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, 2021, pp. 1760–1770.

[24] A. Idrees, Z. Iqbal, and M. Ishfaq, "An efficient indoor navigation technique to find optimal route for blinds using qr codes," in *2015 IEEE 10th Conference on Industrial Electronics and Applications (ICIEA)*, 2015, pp. 690–695.

[25] A. L. Fraga, X. Yu, W.-J. Yi, and J. Saniie, "Indoor navigation system for visually impaired people using computer vision," in *2022 IEEE International Conference on Electro Information Technology (eIT)*, 2022, pp. 257–260.

[26] M. Kuribayashi, T. Ishihara, D. Sato, J. Vongkullbhasal, K. Ram, S. Kayukawa, H. Takagi, S. Morishima, and C. Asakawa, "Pathfinder: Designing a map-less navigation system for blind people in unfamiliar buildings," in *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, ser. CHI '23. New York, NY, USA: Association for Computing Machinery, 2023, pp. 1–16.

[27] N. Shazeer, "Glu variants improve transformer," 2020. [Online]. Available: https://arxiv.org/abs/2002.05202