Real-Time AI-Aided Navigation in Robotic Guidance for the Visually Impaired

Riana Chatterjee, Computer Science
Mentor: Dr. Ransalu Senanayake
School of Computing and Augmented Intelligence

LENS Lab

Abstract

Over 253 million people worldwide live with visual impairments, yet most assistive devices lack customizable navigation capabilities.

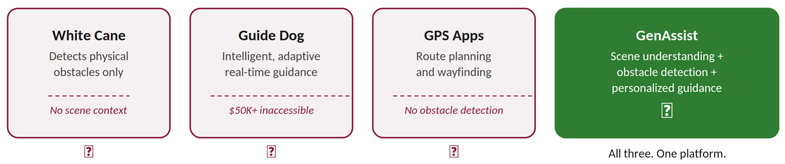
This research presents GenAssist: a framework for quadruped robot navigation using multi-sensor fusion of RGB and LiDAR sensors, integrated with YOLOv8 object detection, Vision-Language Models (VLMs), and Bayesian Optimization to determine optimal, user-specific navigation thresholds.

The system enables the Unitree Go2 robotic guide dog to interpret scenic environments and deliver tailored routing instructions, improving assistive robotic navigation safety across four simulated visual impairment conditions: Age-Related Macular Degeneration, Diabetic Retinopathy, Retinitis Pigmentosa, and Glaucoma.

Introduction

Independent navigation is fundamental to daily life, yet existing assistive tools each address only a narrow part of the challenge. White canes detect obstacles but provide no semantic context. Guide dogs offer intelligent guidance but cost over $50,000 and take 2+ years to train. GPS apps enable route planning but fail to detect real-time obstacles.

A visit to the Arizona School for the Blind in spring 2025 shaped this project directly. Students highlighted the danger of sensory overload and emphasized the need for specific directional language over continuous description. These insights motivated GenAssist, a robotic guide dog built on the Unitree Go2 platform, fusing a RealSense RGB-D camera with YOLOv8 and a locally hosted Vision Language Model to deliver sparse, purposeful guidance.



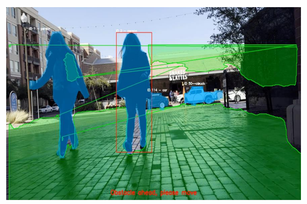


Figure 1. YOLO Inference on camera data

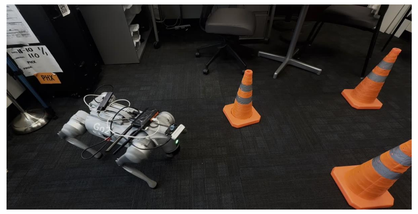


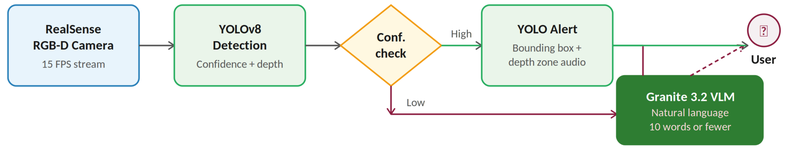
Figure 2. Robotic dog navigating obstacles

Methods and Materials

The GenAssist pipeline streams 15 FPS from an Intel RealSense RGB-D camera into YOLOv8, which assigns each detection a confidence score and real-world depth value. Five color-coded proximity zones classify objects from critical (under 2 feet) to safe (over 5 feet).

When YOLO confidence exceeds the active threshold, the system responds autonomously using bounding box labels and depth. When confidence drops below it, Granite 3.2 generates a natural language description in ten words or fewer, such as "chair two feet to your right." Granite 3.2 was selected after benchmarking eight VLMs on descriptive quality and inference speed.

The Go2 navigates via its ROS2 autonomy stack and LiDAR-based SLAM while the YOLO-VLM layer monitors for hazards in parallel.



Results

At the converged threshold, the VLM was invoked on approximately 1 to 2% of all navigation steps, meaning the robot completed most of its path under YOLO-only detection and paused for VLM inference only on genuinely ambiguous frames. VLM latency averaged 10 to 13 seconds per call on the Jetson hardware, making sparse invocation essential to maintaining navigation flow. User scores at the converged threshold reached 10/10 across multiple sessions. High threshold configurations operating on YOLO detection alone produced scores of 1 to 2/10 universally, confirming that semantic language output is necessary for safe guidance. The fine-tuned YOLOv8 segmentation model achieved 99.5% mAP on sidewalk classification and 97.4% mAP on road classification.

Trial	YOLO Conf at Trigger	YOLO Misclassified Object	VLM Description (truncated)
Live Trial 1 (x=0.510)	0.335	Vase (traffic cone present)	"A bright orange traffic cone placed on the floor. The cone is the most prominent obstacle in the scene."
Live Trial 2 (x=0.520)	0.283	Clock (room obstacle present)	"The room appears to be a simple space with a clear floor. Path appears safe to navigate."
Live Trial 3 (x=0.565)	0.320	None detected (conf below threshold)	"Several potential obstacles and hazards a visually impaired person should be aware of while navigating."
Live Trial 4 (x=0.605)	0.000	None detected (complete YOLO failure)	"Several objects in the room could pose potential obstacles or hazards for a visually impaired person."

Table 1. VLM catch table of when YOLO is incorrect or ambiguous

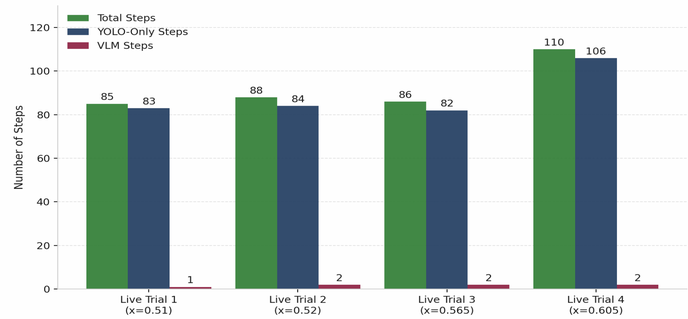


Chart 1. Comparison of Total Steps, YOLO-operated Steps, and VLM operated steps

Discussion

The core finding of this work is that the switch-off architecture succeeds precisely because the VLM is invoked rarely. A system that describes every frame occupies the audio channel that visually impaired users depend on for ambient awareness. The confidence threshold is what makes sparsity possible: YOLO handles the routine frames autonomously, and the VLM steps in only when the scene genuinely exceeds YOLO's ability to characterize it.

The depth integration from the RealSense camera resolved a key early limitation. MiDaS depth-sensors flagged objects by pixel position rather than actual distance, producing false warnings for distant centered objects and missing nearby off-center ones. Per-pixel depth lookup from YOLO bounding box centers replaced this with real-world distance measurements in meters, grounding every alert in physical space.

Conclusions

GenAssist demonstrates that a quadruped robotic guide dog can autonomously navigate indoor environments while delivering sparse, semantically grounded guidance through a confidence-based YOLO-VLM switch-off architecture. The system runs entirely on-device with no cloud dependency, achieves stable 15 FPS perception, and produces user scores of up to 10/10 when the switch-off threshold is correctly calibrated.
The most promising direction for future work is haptic feedback through the leash handle. Currently, all user communication is audio-based. Integrating vibration patterns into the physical leash connection between user and robot would enable directional guidance without occupying the audio channel at all, a direct response to the sensory overload concern raised at the Arizona School for the Blind. Left and right turns, stops, and urgency levels could each map to distinct vibration signatures, making the system usable in loud environments and reducing cognitive load for the user.

FURI

Ira A. Fulton Schools of Engineering
Arizona State University