VLM-Based Navigation for Autonomous Mobile Robot

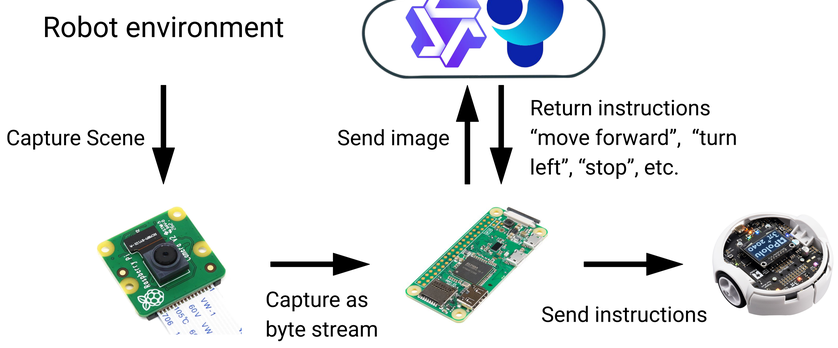
Berkeley | EECS
Electrical Engineering and Computer Sciences
EECS 149/249A - 2025

Robin Rheem, Jeffrey Zhang, Kushagra Srivastava

Motivation

- Traditional navigation relies on **geometry only** (e.g. point clouds, ray casting)
  - Good for collision avoidance, but lacks **semantic understanding**
- Vision-Language Models (VLMs) add high-level scene reasoning
- Goal: use VLMs to help robot choose best navigable path (largest visible gap between obstacles), then navigate through it

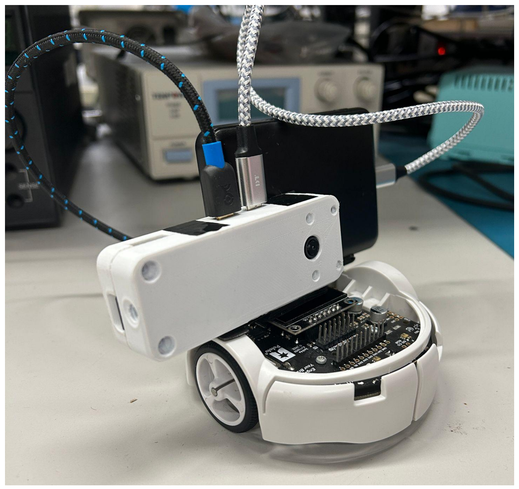
System Design



The navigation pipeline operates as follows:

- Onboard camera captures images of the environment.
- Images are sent to a remote server for processing.
- VLM identifies relevant objects in the scene.
- Object detection model produces bounding boxes and spatial coordinates.
- Geometric reasoning determines the optimal travel direction.
- The robot receives and executes the resulting motion command.

This loop repeats until the robot successfully navigates through the largest detected gap.



Implementation Overview

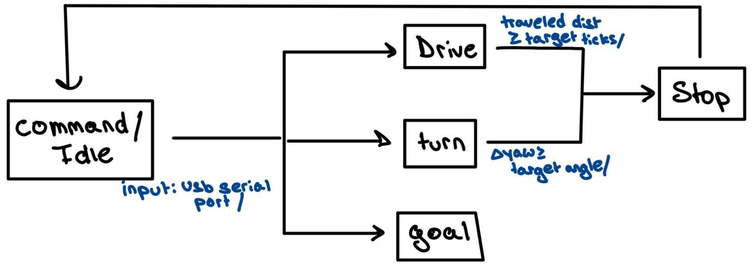
Hardware

- Raspberry Pi Zero W (platform for camera module, wireless connection w/ server)
- Raspberry Pi Camera Module 2
- Pololu 3pi+ 2040 differential drive robot
- Inference Server (24 core, 64 GB RAM, No GPU)

Software

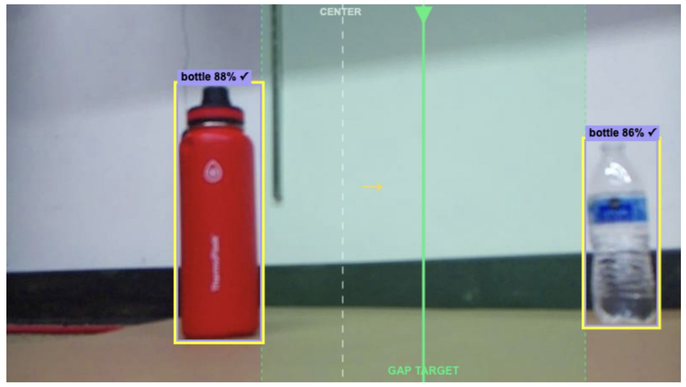
- Qwen2.5-VL-7B-Instruct - visual and spatial reasoning for understanding robot’s environment
- YOLOv11 - real-time object detection system using CNNs
- Python for RPi Camera, Inference Infrastructure
- Lingua Franca and C for Pololu robot

Robot State Machine



Results

- VLM-only navigation provides semantic understanding but lacks precise spatial grounding
- **accurate navigation requires both semantic reasoning *and* geometric localization**
- Hybrid System (**VLM + YOLO**) combines high-level scene reasoning with reliable object positions
- Outcome: robot consistently identifies and navigates largest visible gap for indoor controlled environments



Results of navigation system

Room for Improvement

- Camera Latency:
  - Pi Zero W has slow camera warm-up & capture
- Inference Latency:
  - CPU-only Inference increases response time
- Model Accuracy
  - Limited object vocabulary and labeling completeness
- Environment Generalization
  - Performance may degrade in cluttered or noisy real-world environments

Class Concepts

- Sensors/Actuators
- State Machines
- TinyML/Wireless Communication