ATLAS

Adaptive Technology for Location and Spatial Support

Problem Statement

The Problem

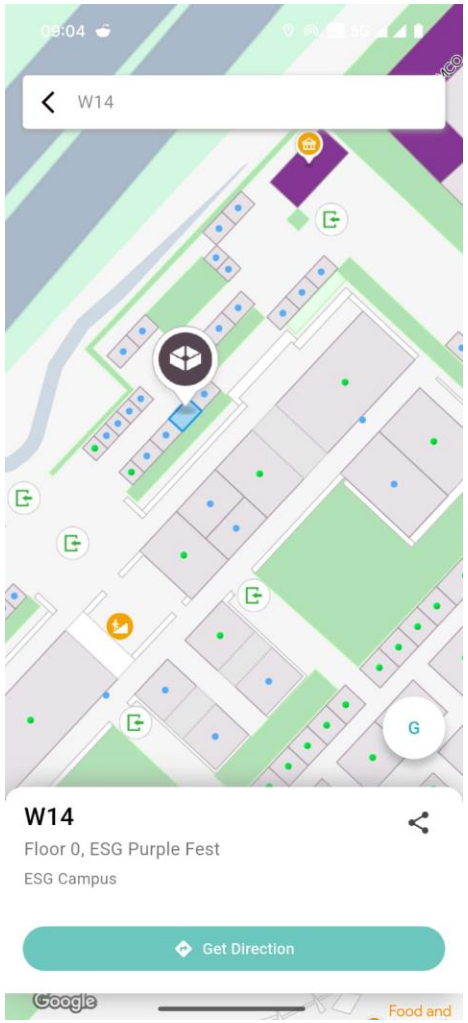
- Individuals who are visually impaired require an **escort to navigate unfamiliar spaces**
- **Outdoor and crowded environments** are harder to navigate without a companion
- Existing tech solutions either **lack precision or are expensive**

How can we **enhance navigation systems** for persons with visual impairment to make **information** and **movement** more **accessible**?

Stakeholder Insights

Existing Tech

- **Google Maps**
  - Need **screen reader** to know when distance hits zero for turns. Need to hold out phone.
  - GPS doesn't perform well in **indoor scenarios, or India.**
- **iWayplus (IIT Delhi, Prof. Bala)**
  - Indoor Navigation System using **beacons**, attempting to improve over GPS. More precise, but manual set-up required.
  - **Does not account for changing environment** variables like people or obstacles



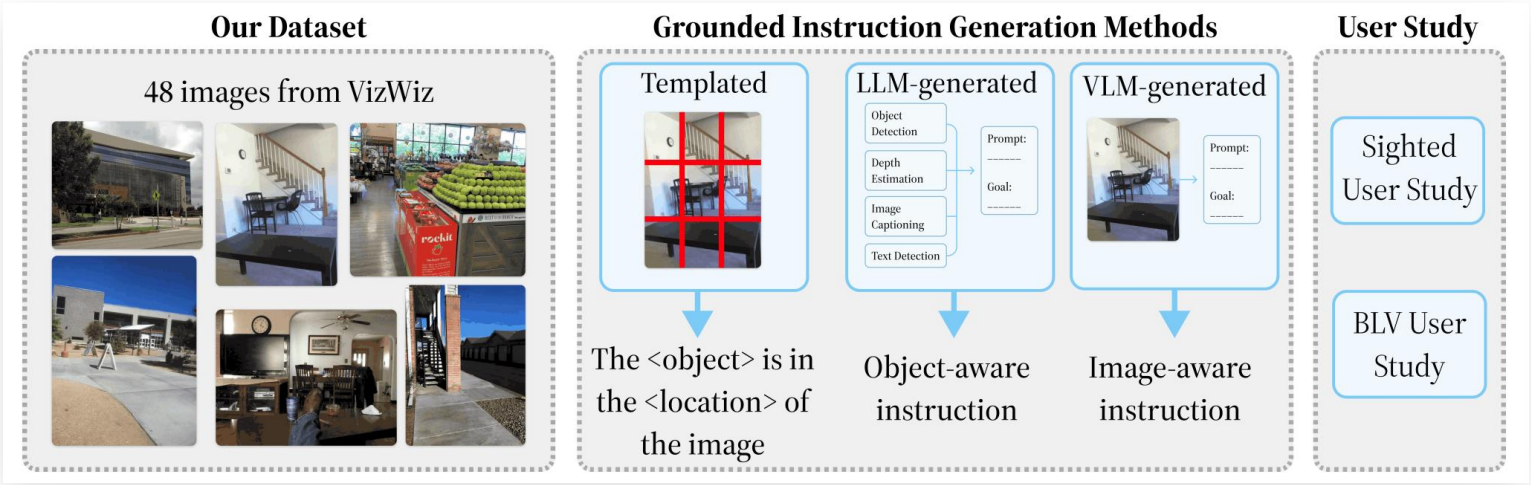
iWayPlus in action

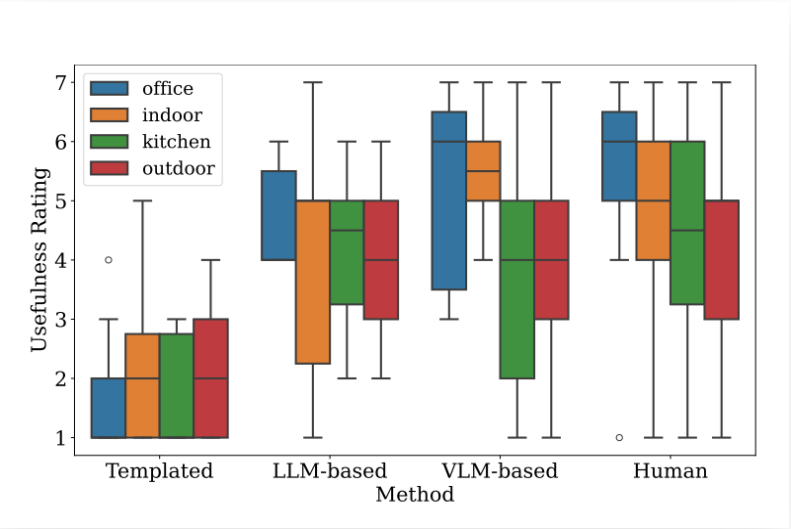
Existing Tech

- **Meta Raybans (Smart Glasses)**
  - **Good microphones** for input sensitivity
  - **Good for Q&A** and scene description via native apps like BeMyEyes
  - Requires **Phone + Internet Connection**
  - **High Latency (30s-1min)** is not ideal for real-time feedback situations
  - Glasses with haptics **start paining** and are **not suitable for prolonged use**.

Literature Review

Generating Contextually Relevant Navigation Instructions for Blind and Low Vision People

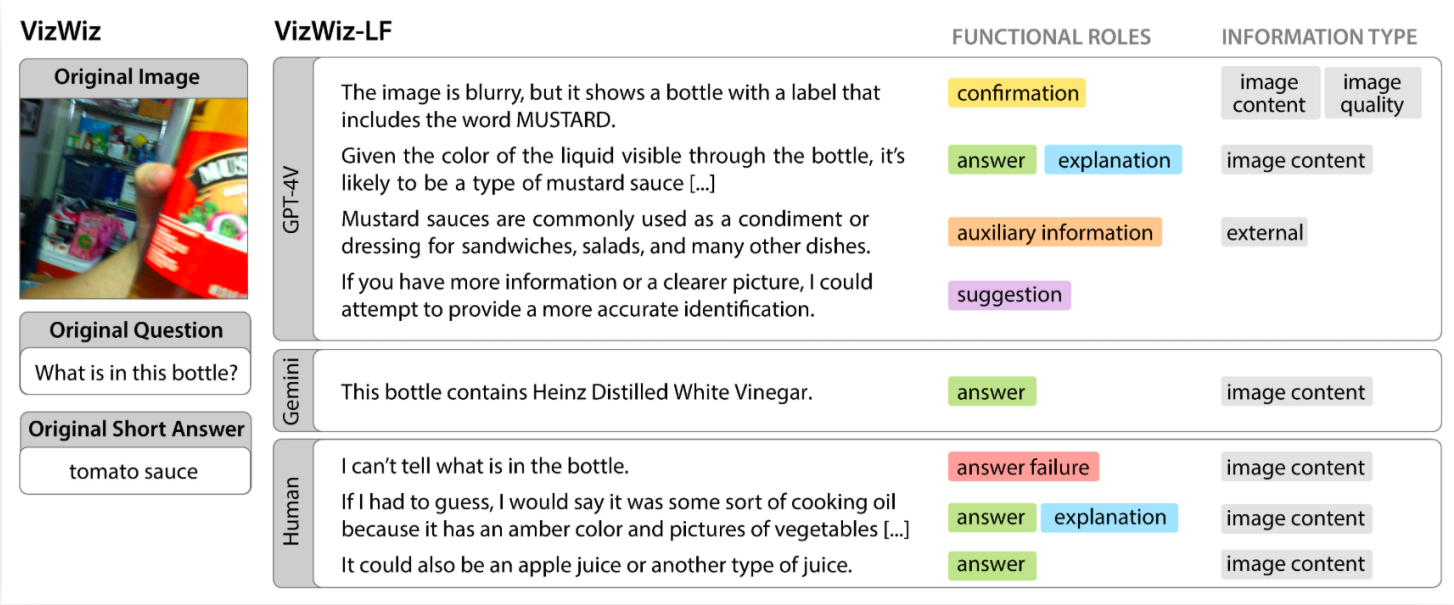




- Sighted and BLV user study
- 3 different methods tested — **OWL-ViT** (Templating) vs **GPT-4** (LLM) vs **GPT-4 Vision** (VLM)
- Verbose and **rule-based prompt conditioning**
- Found it challenging due to **frequent hallucinations** which in some instances is dangerous

Long-Form Answers to Visual Questions from Blind and Low Vision People

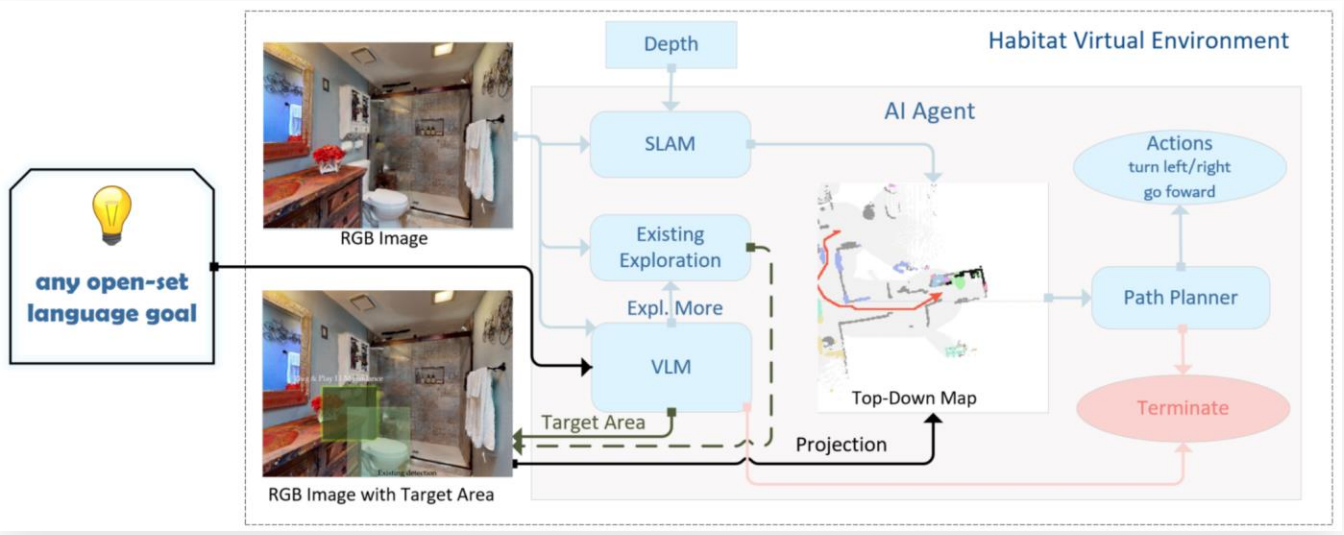
- **4.2k long-form answers** to 600 visual questions
- Prevent hallucinations from unanswerable questions
- Human experts vs VLMs
- 20 BLV and 20 sighted users rated relevance, helpfulness, plausibility/correctness, and fluency
- **GPT-4V scored the highest** amongst BLV users — they prefer longer answers in certain scenarios



Huh, Mina, et al. (2024) - Conference on language modelling

Navigation with VLM Framework: Go to Any Language

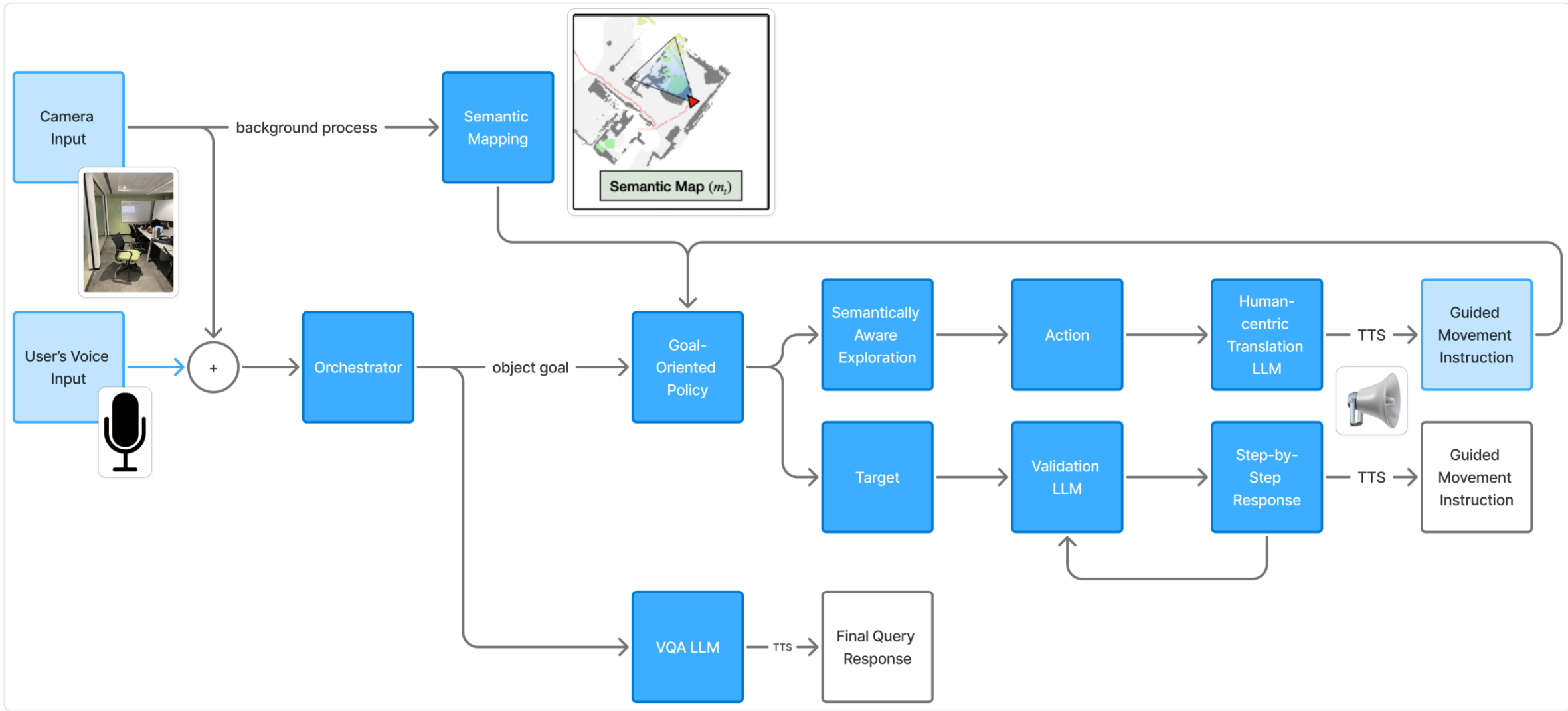
- **Zero cost adaptability** on unknown environments
- VLM core runs on **mobile device**
  - Lower costs and lower latency
- Dataset: **Habitat Matterport3D**
- Support for **ambiguous language goals** such as "somewhere I can sit and eat"
- Best generalization when VLM used for **rough location** instead of end to end or precise location
- Average SPL Score – **39.2**



Framework for modelling the system in HabitatSim

Yin, Zecheng, and Chonghao Cheng 2024 - In review arXiv preprint

Methodology

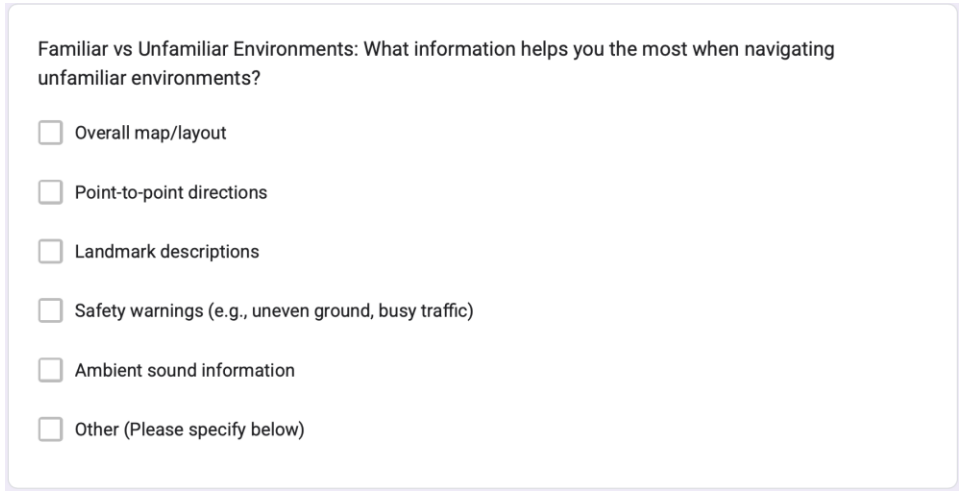

Camera Input
background process
Semantic Mapping
Semantic Map (m_t)
User's Voice Input
+
Orchestrator
object goal
Goal-Oriented Policy
Semantically Aware Exploration
Action
Human-centric Translation LLM
TTS
Guided Movement Instruction
Target
Validation LLM
Step-by-Step Response
TTS
Guided Movement Instruction
VQA LLM
TTS
Final Query Response

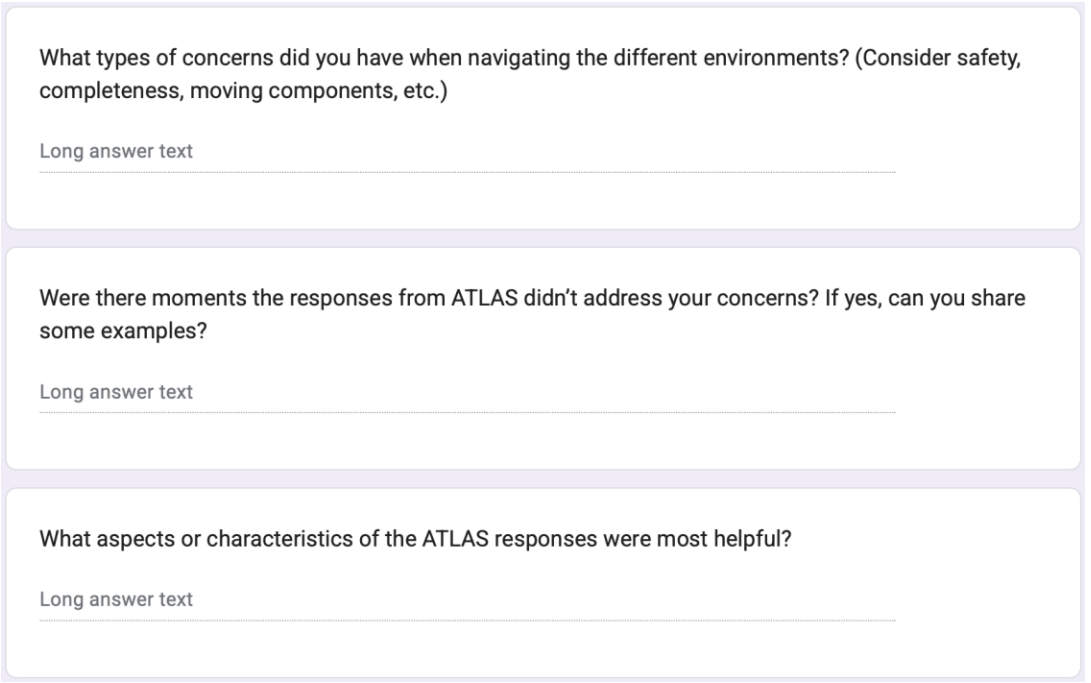
Evaluation Metrics

- **Objective Metrics**
  - VQA - Accuracy calculated based on LLM evaluation
  - Navigation - Success weighted by path length (SPL)
  - Response Latency (ms)
  - $\mathrm{SPL} = S \cdot \frac{P_t}{P_a}$
  - $(where\ \mathrm{P_t} = \text{taken path length and},\ \mathrm{P_a} = \text{best path length})$

Subjective Evaluation

- 8 individuals, 4 sighted and 4 with blind folds.
- Evaluated user-feedback post survey with the questions like:

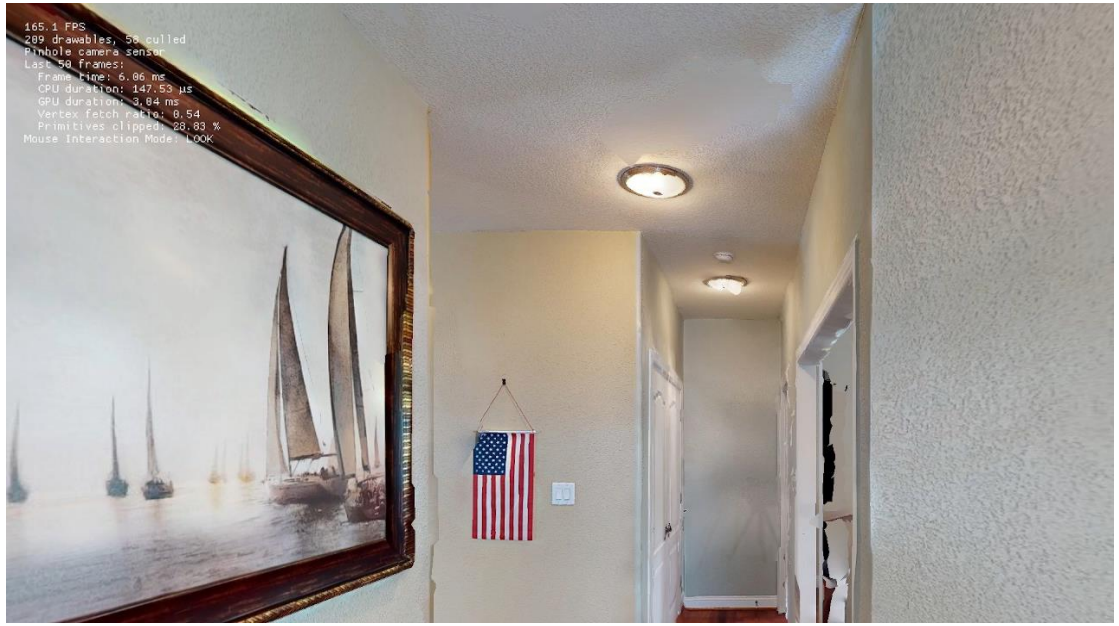




Analysis and Feedback

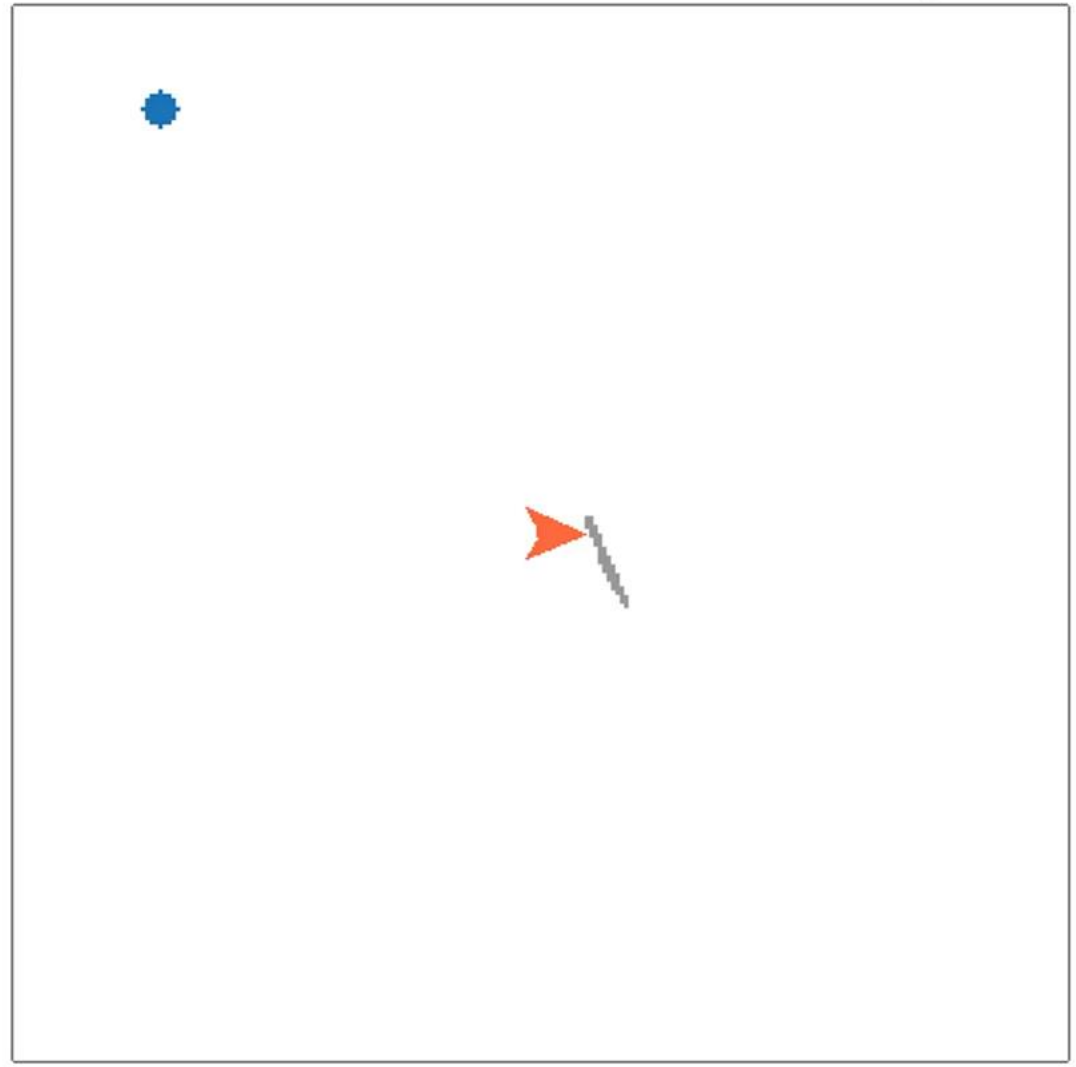
Dataset – Habitat Sim

- **HM3D**: largest ever dataset of 3D Indoor spaces
- Standard simulation space for **embodied AI**
- **Nondeterministic non sighted individual human controller**
- Allows for us to run **multiple trials**

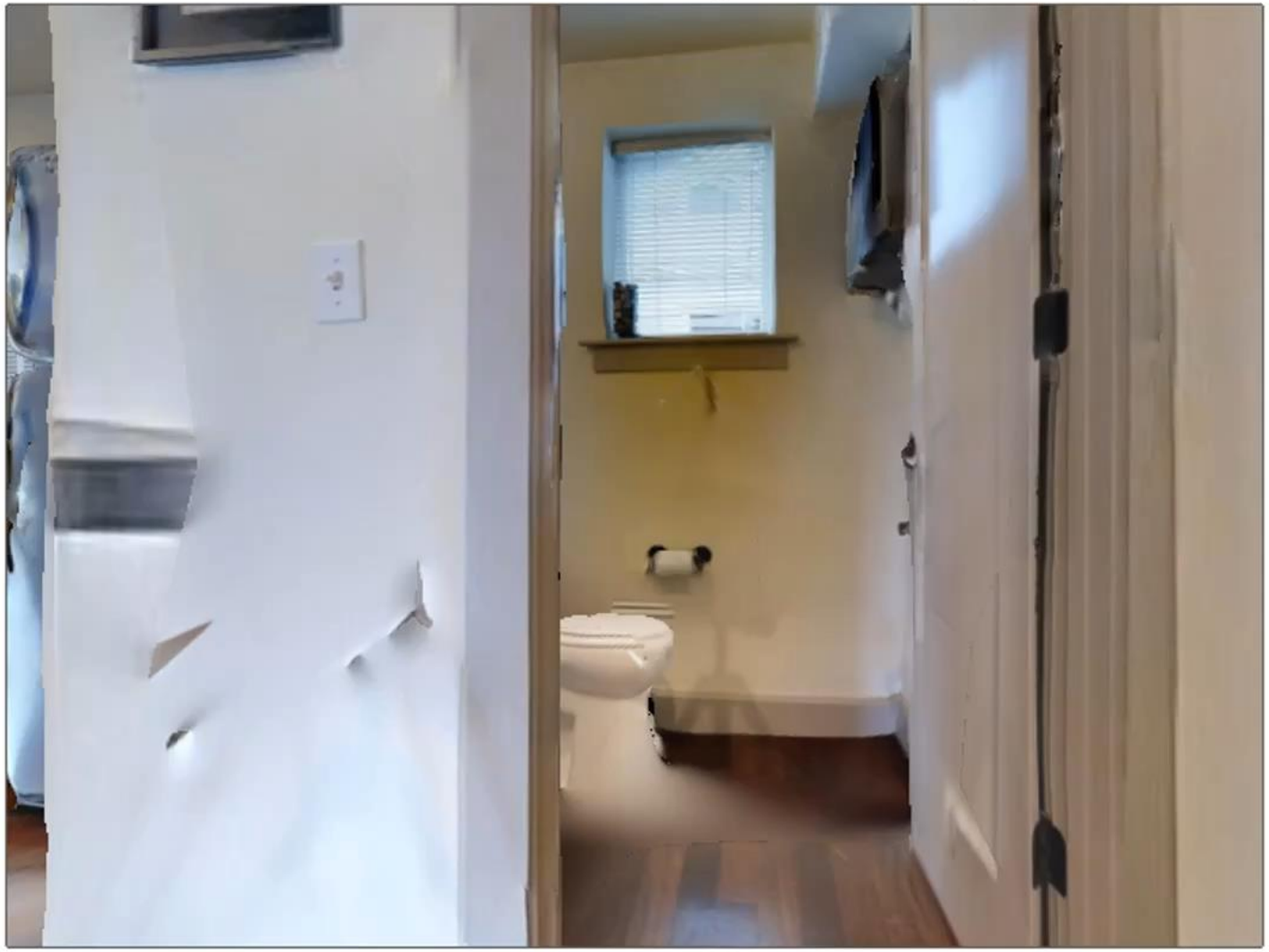


Observations (Goal: toilet)

Predicted Semantic Map

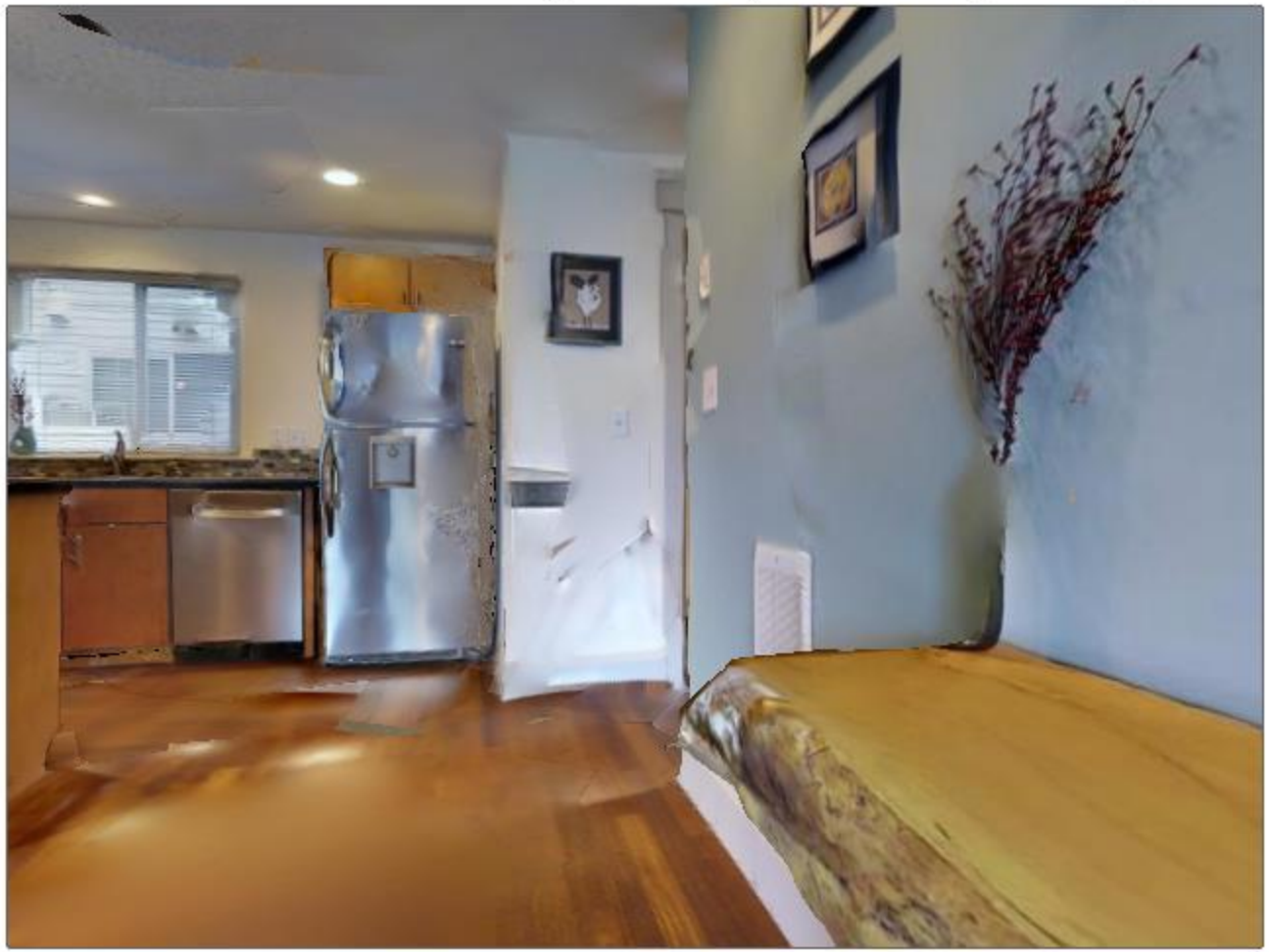


Navigable Area
0: chair
1: couch
2: potted plant
3: bed
4: toilet
5: tv
6: dining-table
7: oven
8: sink
9: refrigerator
10: book
11: clock
12: vase
13: cup
14: bottle

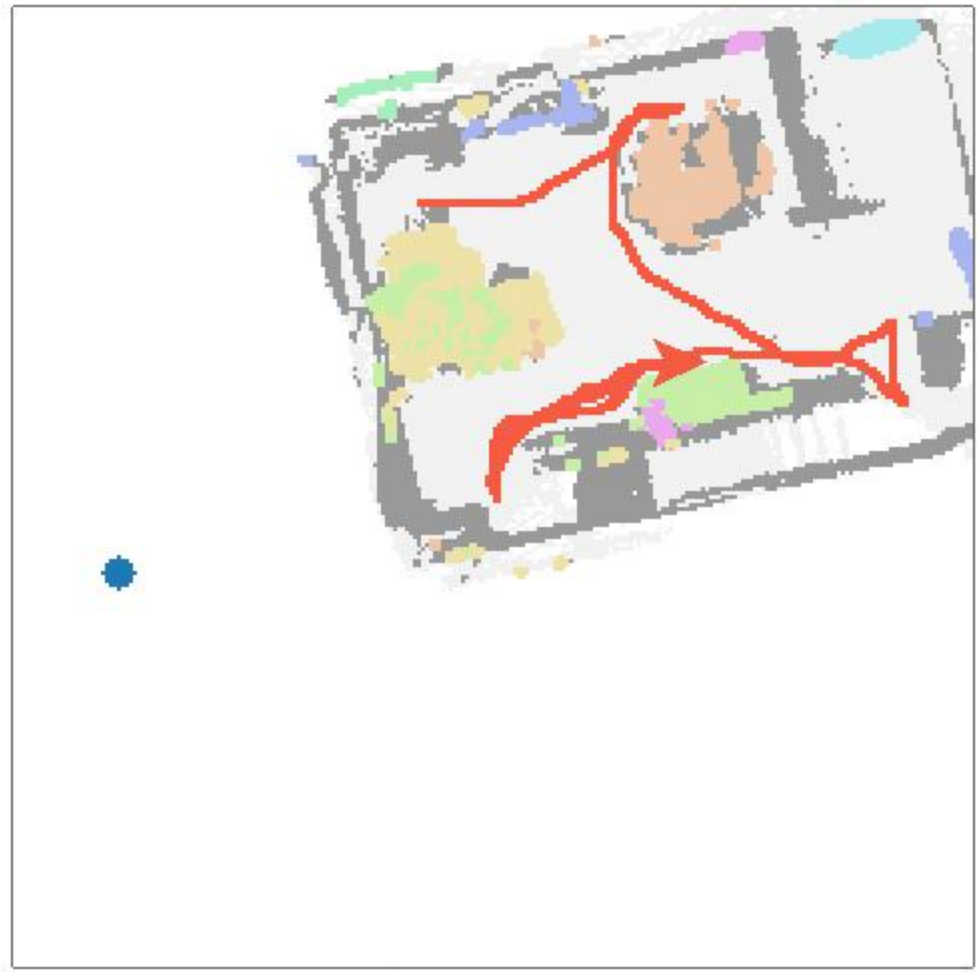


Navigable Area
0: chair
1: couch
2: potted plant
3: bed
4: toilet
5: tv
6: dining-table
7: oven
8: sink
9: refrigerator
10: book
11: clock
12: vase
13: cup
14: bottle

Observations (Goal: potted plant)



Predicted Semantic Map



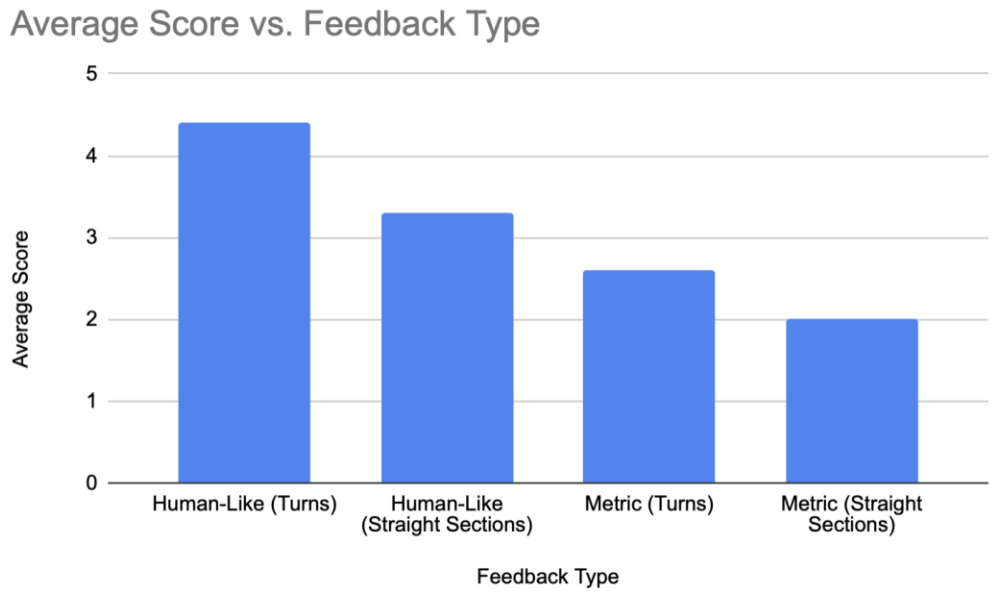
Navigable Area
0: chair
1: couch
2: potted plant
3: bed
4: toilet
5: tv
6: dining-table
7: oven
8: sink
9: refrigerator
10: book
11: clock
12: vase
13: cup
14: bottle

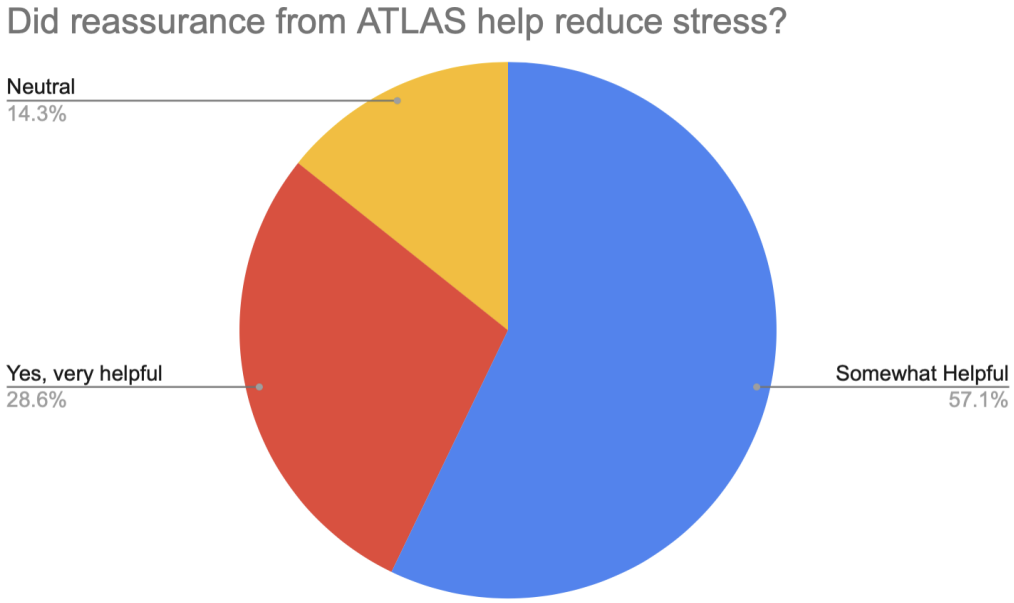
Sim Results

Goal	Success Rate	SPL
Toilet	1.0	0.829
Chair	0.75	0.724
TV	1.0	0.945
Potted Plant	0.25	0.145

Each Goal was tested for 20 episodes

Post-Survey Results






Amul
Lassi

VQA Accuracy

Prompt Type	Evaluation Type	LLM Evaluation using Gemini 2.5 Pro
Description	Relevance Score	83.5%
Detection	Object Detection	90%
Information Extraction	Correctness	80%

20 question-image pairs were tested for each type.

Deployability

Deployment Considerations

ATLAS Module	Off the Shelf Component
RGB-D Camera	iPhone w Lidar, DepthPro Monocular Depth Estimation, Intel RealSense Cameras
SLAM Inference	SemExp achieves 15 FPS on A6000
Device Platform	Meta RayBans, mobile phones
VLM for VQA System	for Gemini
Text-to-Speech System	Realtime performance on-device using PyTTSX3

Impact

ATLAS IMPACT

- **35 million** people in India suffer from some vision ailment (*Indian journal of ophthalmology*, 2022)
- Uses context and semantic aware search for locating goals in **unknown** / **occluded environments**
- BeMyEyes requires a **live connection to a human** for guidance in unknown environments
- Zero-shot **multilingual** adaptability of the pipeline
- Supports Open-set Language Goal: Extends the language goal set from traditional specific goals to a fully **open language set**

Challenges

Challenges

- **Reaching out to the Institute For Blind –** No response
- **Lack of low vision human walking model**
- **Making the system proactive**
  - Frequently updating context with live camera input
  - Need Large Context window models
- **System working in all Network Conditions seamlessly**
  - Create a fallback mechanism using a local LLM that can perform basic tasks

thank you.