Conversational Navigation for Smart Glasses

Mixed Reality - Fall 2025

R.Bianco, F.Bondi, R.Nobari, S.Panwar, F.Daneshmand

Supervised by: M.Rad, G.Goletto, K.Jaroslavceva

ETH zürich

Microsoft

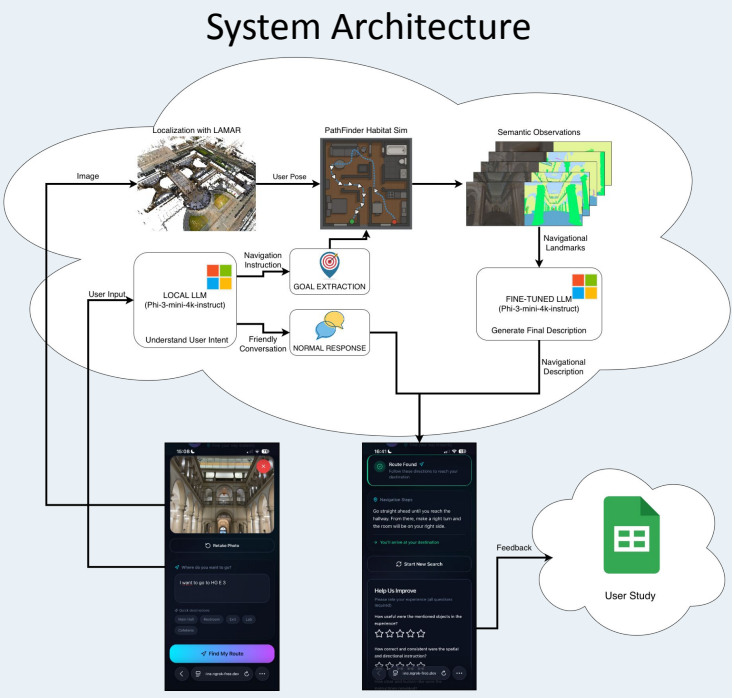
1 Introduction

In this work, we built a **natural-language navigation system** inside the **HG building (floor E)**, using Habitat-Sim as a simulation backend. From any starting point, users can ask questions such as "How do I get to HG E 3?", and our pipeline simulates a path, gathers visual observations, and produces **human-like navigation instructions** grounded in objects and scene cues. This work explores how embodied agents can integrate simulation, vision, and language to generate intuitive indoor navigation guidance.

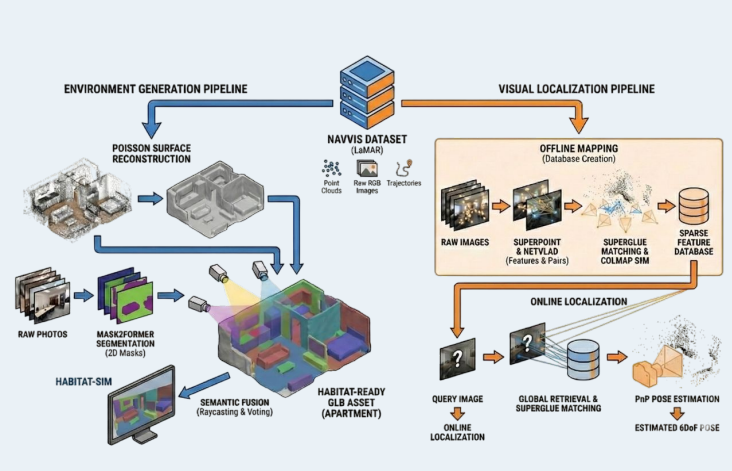
2 Background

Vision-and-Language Navigation (VLN) studies how agents follow natural language instructions while moving in realistic environments. Earlier methods operated on discrete navigation graphs, whereas VLN-CE introduced continuous, low-level navigation with egocentric perception, making the task more realistic and challenging. Recent advances in **multimodal Large Language Models** (Video-LLMs) and instruction tuning, such as LLaMA-VID, have improved the grounding of visual observations and language. Models like StreamVLN further show that combining Video-LLMs with efficient long-term memory enables **coherent navigation over continuous video streams**. These developments point toward **increasingly capable embodied agents**, forming the foundation for our natural-language navigation system in Habitat-Sim.

3 Method

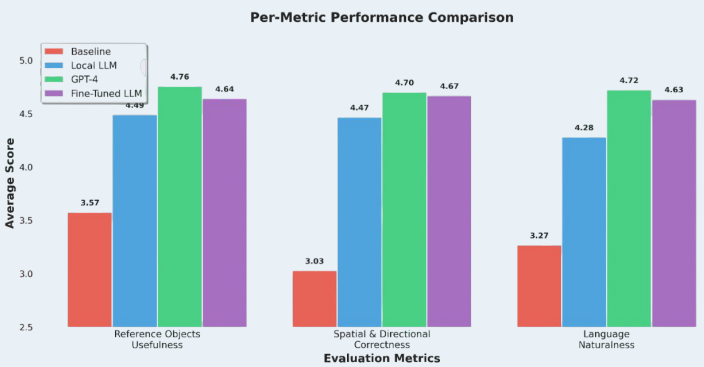


- **Web App**
  Send a picture from your current position, include a text and send it to the server.
- **Intent Understanding (Local LLM)**
  Interpret the user's natural-language request and extract the navigation goal.
- **User Localization (LAMAR)**
  Estimate the user's 6-DoF pose in the real environment.
- **Path Planning (Habitat PathFinder)**
  Compute a navigable path from the user's position to the target location.
- **Semantic Observation Extraction**
  Collect visual and semantic landmarks along the planned path.
- **Route Description Generation (Fine-Tuned LLM)**
  Convert paths and semantic cues into step-by-step natural-language instructions.
- **User Feedback Collection**
  Gather qualitative feedback to evaluate usability, accuracy, and perceived performance.



4 Results

We evaluate navigation instructions using a user study across three objective dimensions: **reference objects**, **spatial correctness**, and **language naturalness.**



Key Takeaways

- **Local LLMs are effective for navigation instruction generation**
  Even without fine-tuning, a compact local model (Phi-3-mini-4k-instruct, 4B parameters) substantially improves navigation quality over the baseline across all evaluation metrics.
- **Fine-tuning bridges the gap with large proprietary models**
  A fine-tuned Phi-3-mini model achieves performance comparable to GPT-based solutions, particularly in spatial correctness and language naturalness.
- **Strong performance with orders-of-magnitude fewer parameters**
  Comparable navigation quality is obtained using a 4B-parameter local model, despite a large gap in model scale with proprietary LLMs.
- **Implications for deployable navigation systems**
  These results suggest that high-quality natural-language navigation can be achieved with lightweight, deployable models, enabling on-device applications.

5 Conclusions & Future Directions

This work demonstrates that **natural-language indoor navigation** can be effectively generated by combining **path planning**, **semantic observations**, and **lightweight language models**. Our pipeline significantly improves navigation quality over a baseline approach, achieving strong performance across spatial correctness, reference object usage, and language naturalness. Notably, a compact local LLM (Phi-3-mini, 4B parameters) already provides substantial gains, while fine-tuning enables performance **comparable to large proprietary models**, highlighting a favorable trade-off between model size and instruction quality. Future work will focus on **real-world deployment** using existing wearable hardware, replacing the current mobile interface with real-time sensory inputs such as egocentric vision, and enabling **fully on-device assistive navigation systems**.