Can GPT-4o Be a Robot's Spatial Reasoner?

Shun Jin, Weison Huang

15-494 Cognitive Robotics | Final Project

Vision-language spatial reasoning on the VEX AIM robot

The Problem

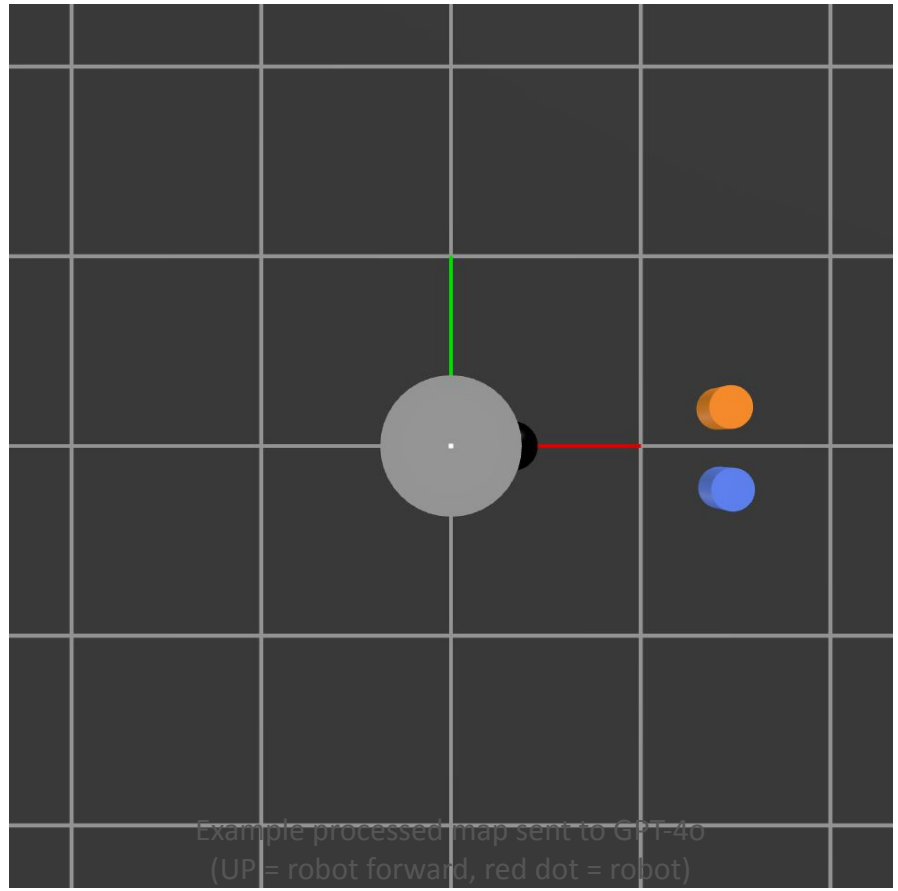
Turning a SLAM world map into language-level understanding and plans

- Traditionally, extracting answers and navigation plans from SLAM world map requires hand-coded geometry: distances, occupancy checks, path planning.
- Can a vision LLM (GPT-4o) reason about the scene directly from a top-down world map image?
- Goal: replace as much hand-coded spatial logic as possible with natural-language prompting.
  - counting
  - relational questions
  - free-space checks
  - multi-step path planning

Our Approach

World map -> preprocess -> GPT-4o -> parse commands -> FSM drives the robot

- Capture the live world-map viewer as a PNG each round.
- Preprocess of the world map image.
- Send image + per-round scale line ('k pixels = 100 mm') to GPT-4o with a structured system prompt.
- Prompt:
    - enforces pixel-measurement -> millimeter conversion
    - output in robot’s body frame.
- Parse #forward / #sideways / #turn / #pilottoobject / #pickup commands plus [EXPLORE] / [EXECUTE] / [DONE] tags.

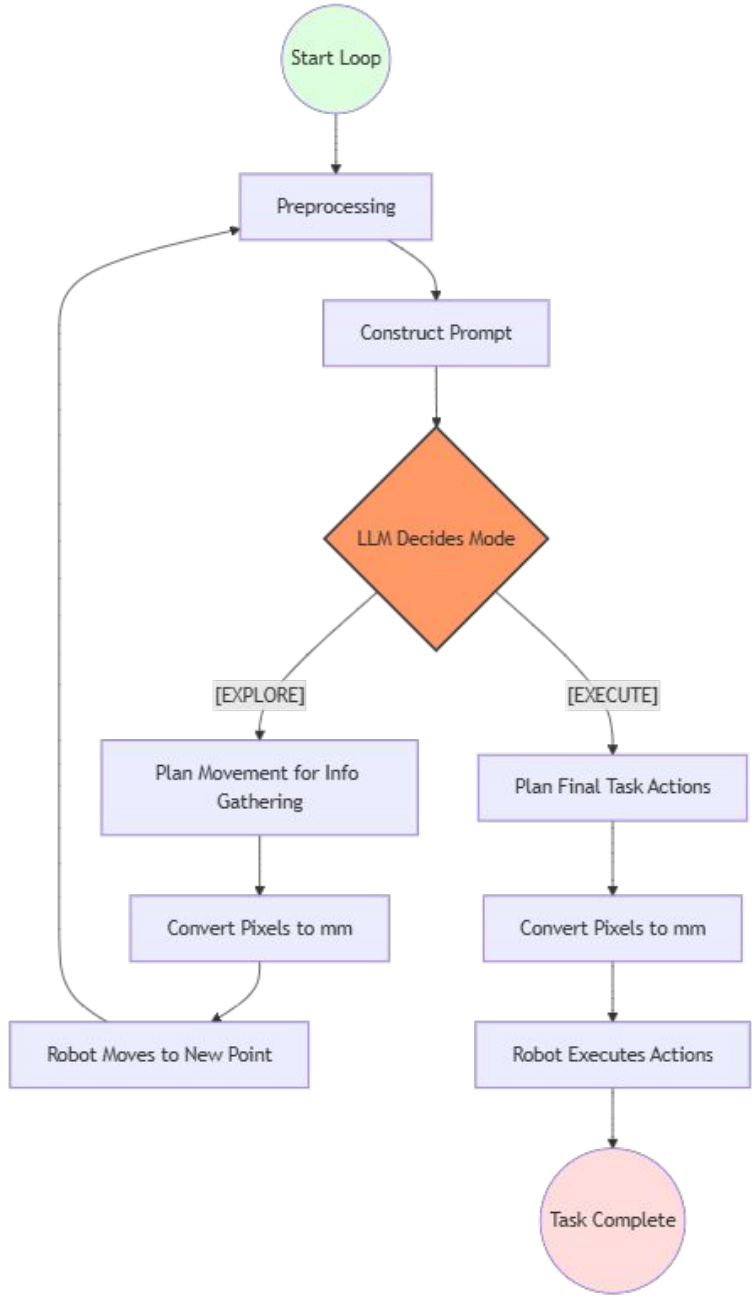


Example processed map sent to GPT-4o
(UP = robot forward, red dot = robot)

Most Interesting Aspect

An LLM-driven explore / execute loop with no hand-coded geometry

- Two modes, chosen by the model itself each round (loop):
  - [EXPLORE] to gather more map info
  - [EXECUTE] to commit to actions.



Results - What Worked

End-to-end pipeline runs on the real robot

- Auto world map capturing
- GPT-4o produces step-by-step spatial reasoning in first-person.
- Exploration / execution loop works end-to-end:
  - the robot can explore, re-query, and then act on short tasks.
- Simple question-answering
  - ('what's around me?', 'what's ahead?') is generally reasonable.

Results - What Didn't Work

Where GPT-4o's spatial reasoning breaks down

- Distance estimates drift:
  - GPT-4o often mis-measures pixel offsets
- Error in orientation estimation
  - Despite pre-rotating the image so UP = forward
- Hitting obstacles when world map’s objects are close

Future Work

Closing the gap between LLM reasoning and metric geometry

- Pass structured object metadata (type, position, color) as text alongside the image.
- Give GPT-4o tool use: a 'measure(object)' or 'free_space(direction)' function it can call instead of guessing.
- Add change-detection across successive maps to support dynamic-environment tasks.
- Try stronger models in image analysis (e.g. Gemini 3)


AIM-526DA01S