LLMs as Assistive Visual Information Parsers for the Visually Impaired

Motivation: We explore the cognitive capabilities of large language models (LLMs) to reason about a scene based solely on knowledge of relevant objects and their spatial positions. We designed a specialized module that leverages these capabilities to assist visually impaired individuals in navigating and interpreting their environments.

Literature Survey: For our work, we reviewed literature in the fields of 3D spatial reasoning (e.g., 3D LLMs[1])) and language-image paired learning (e.g., BLIP[2], CLIP[3]). Since our approach specifically integrates object detectors with LLMs, we also studied object detection methods, focusing particularly on YOLO[4].

Problem Definition: We formalize our problem statement as: "Given a stream of RGB-D video, how can we process this information using an LLM to generate coherent language output that informs a visually impaired person about their surroundings?"

Proposed Idea: Our methodology uses an object detector to first identify and localize objects in the scene. These detections are then paired with depth information and provided to the LLM for reasoning. This approach allows the LLM to incorporate temporal data, enabling better understanding of the scene. Also, as object detectors are computationally lighter than VLMs, our method is less resource intensive. The pipeline for our work is illustrated in Fig. 1.

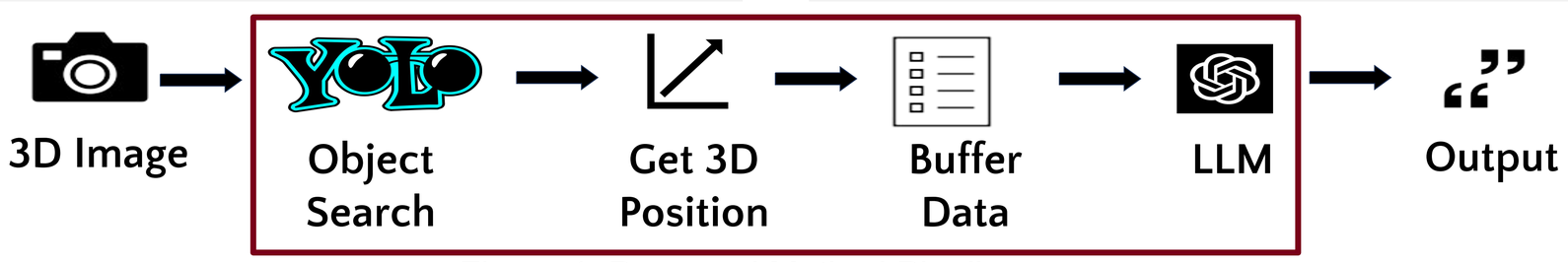

Figure 1. Proposed Method Pipeline

Method: We envision our product to have separate modules made for specific tasks that could be encounter in daily life of the user.

1 Navigation

- This module help the user in navigating their environment.
- Images captured at regular intervals by a LiDAR camera (Intel RealSense L515) are passed through an object detector to extract object labels and pixel locations. The pixel locations, combined with the depth information, are used to calculate the 3D spatial positions of the objects.
- The 3D spatial data is parsed using a proximity map we developed, which translates numerical information into textual descriptions to enhance the reasoning capabilities of the LLM. The LLM is then queried with the last 15 seconds of data and the outputs of the previous five LLM responses to generate a coherent, user-friendly description.
- Specifically, we use YoloV11 for Object detection & GPT 4o-mini as LLM.
- Methodology of this module is shown in fig. 2.

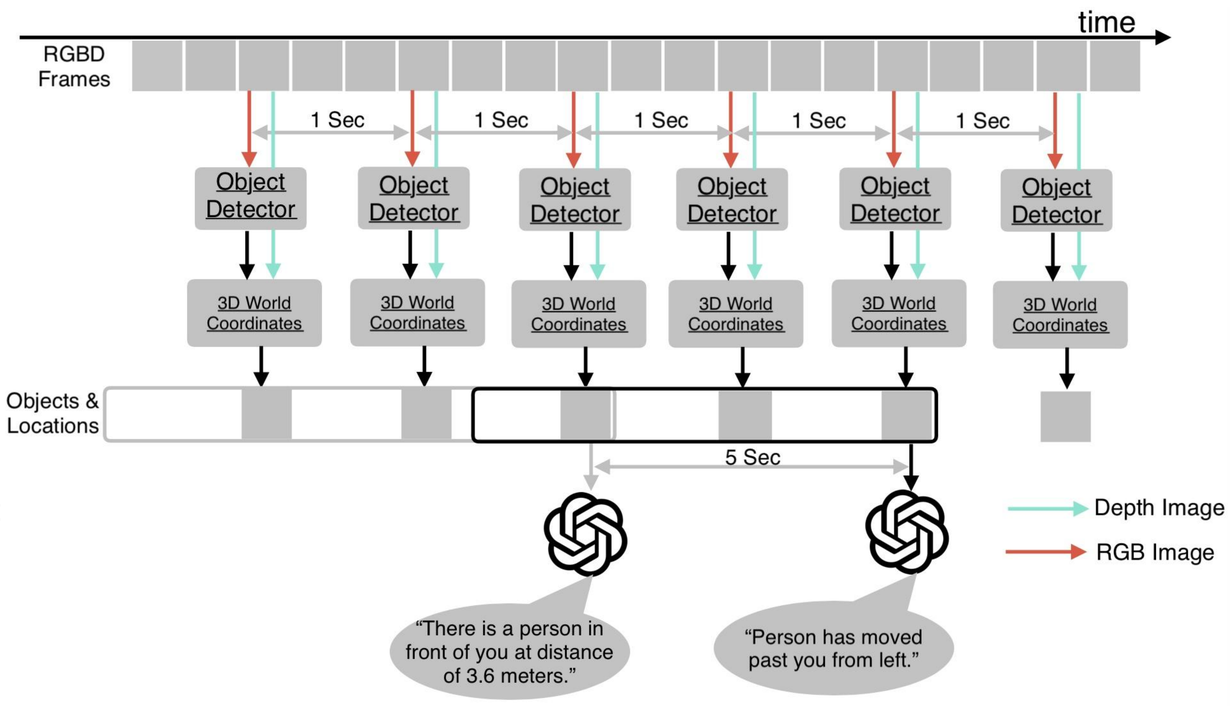

Figure 2. Navigation Module Pipeline

2 Visual Question Answering

- This module helps the user explore the static scenes more in depth.
- We use an off the shelf Visual Question answering package (BLIP) for providing response to the user's multiple queries.

3 Scene Description

- This module uses a off the shelf pretrained BLIP model to Generate a caption for the image. For use cases when user is interested in a general description of the scene.

Experimental Results and Comparison: We have a live demo of our work and have evaluated it qualitatively so far. A quantitative comparison with a baseline VLM will be included in the final report.

Main Findings: In our qualitative tests, we found that the LLM demonstrates strong coherence in reasoning about the relevant objects, based on their location and object labels. Additionally, we observed that textual information, rather than numerical data, aids the model in better understanding the scene.

Limitations and discussion: Our method relies on object detection and depth information from the camera, which may not always be accurate. While the approach shows promising results, there is significant potential for improvement in the system design to better leverage the capabilities of the LLM. For instance, incorporating IMU data could help the system better reason about the user's movements.

Plan for the final report: We have not yet conducted a quantitative comparison of our method with a baseline VLM. We plan to collect human evaluation data for this comparison and will report our findings in the final report.

References:
1. Hong, Y., Zhen, H., Chen, P., Zheng, S., Du, Y., Chen, Z., & Gan, C. (2023). 3D-LLM: Injecting the 3D World into Large Language Models.
2. Li, J., Li, D., Xiong, C., & Hoi, S. (2022). BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation.
3. Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). Learning Transferable Visual Models From Natural Language Supervision.
4. Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You Only Look Once: Unified, Real-Time Object Detection.

Project Webpage

Scan to visit the project webpage

Sentimentals
Mohit Yadav, Alex Besch, Abbas Booshehrain, Ruolei Zeng
yadav171@umn.edu, besch040@umn.edu, boosh002@umn.edu, zeng0208@umn.edu
University of Minnesota