

Navigational Assistance for the Blind in Complex Indoor Spaces Using a Vision-Enabled Large Language Model

Bamikole Adewale, Chantal Lemieux, Yue Zhang, Mahreen Nasir, Rashid Hussain Khokhar, Ajmery Sultana and Wenjun Lin*

Abstract This study introduces an innovative implementation of a Large Language Model (LLM) that leverages both vision and natural language processing to enhance navigation for individuals who are blind. Unlike traditional methods that rely on pre-existing maps or environmental reconstruction using sensors like LiDAR, our approach requires no prior environmental data and instead utilizes real-time visual cues similar to human navigation strategies. This novel methodology allows the model to dynamically interpret and verbalize complex indoor environments, providing blind users with descriptive audio cues that effectively convey the spatial layout and pertinent features of their surroundings. Conducted in a hospital setting,

Bamikole Adewale

Digital Healthcare Innovation Lab, School of Computer Science & Tech., Algoma University, Brampton, Canada, e-mail: badewale@algomau.ca

Chantal Lemieux

Psychology Department, Algoma University, Brampton, Canada e-mail: chantal.lemieux@algomau.ca

Yue Zhang

Mathematics and statistics, Thompson Rivers University, Kamloops, BC, Canada e-mail: yuezhang@tru.ca

Mahreen Nasir

School of Computer Science & Tech., Algoma University, Saulte Ste. Marie, ON, Canada e-mail: mahreen.nasir@algomau.ca

Rashid Hussain Khokhar

School of Computer Science & Tech., Algoma University, Saulte Ste. Marie, ON, Canada e-mail: rashid.khokhar@algomau.ca

Ajmery Sultana

School of Computer Science & Tech., Algoma University, Brampton, Canada e-mail: ajmery.sultana@algomau.ca

Wenjun Lin*

Digital Healthcare Innovation, Lab School of Computer Science & Tech., Algoma University, Brampton, Canada e-mail: randy.lin@algomau.ca

our experiments demonstrated that this approach significantly improves GPT4-V’s navigation capabilities and offers real-time, contextually relevant guidance, thereby enhancing the independence and safety of blind individuals navigating complex spaces. This research contributes to the understanding of AI’s capabilities in real-world applications and opens new avenues for the deployment of language models in complex, dynamic environments.

1 Introduction

Navigating complex indoor environments is a significant challenge for individuals who are blind, as traditional navigation aids often fail to adequately address the dynamic and intricate layouts of places such as hospitals. Recognizing these challenges, this research introduces a groundbreaking implementation of a LLM that leverages its capabilities in both vision and natural language processing to enhance autonomy and accessibility. This novel approach aims to mimic human reasoning, providing real-time support and enabling blind individuals to navigate independently and safely.

The core innovation of our approach lies in the LLM’s ability to process visual information and translate it into descriptive navigational cues. The approach expands our existing Adaptive User Interface framework [7] and adds visual cues for navigation. Traditional navigation aids often rely on pre-existing maps or require the environment to be recreated using sensors such as LiDAR [17]. These methods depend heavily on preliminary knowledge of the environment and are primarily limited by focusing solely on the 3D structure of the space. While LiDAR can accurately map physical layouts and detect obstacles, it fails to interpret non-structural elements crucial for contextual navigation, such as distinguishing between different types of spaces or recognizing essential signs and symbols. This limitation becomes particularly pronounced in dynamic settings like hospitals, where understanding the function of a space or adapting to temporary changes in layout is necessary for effective navigation. In contrast, our proposed GPT4-V based method, **Structured Data Capture (SDC)**, does not require any prior mapping of the space and utilizes visual cues, similar to those used by humans, to interpret and describe the surroundings in real-time. This allows the model to dynamically adapt to environmental changes and provide immediate, contextually relevant guidance to blind users, significantly enhancing their ability to navigate complex indoor spaces independently.

The implications of this research enhance the navigation capabilities of individuals who are blind, and significantly improves the autonomy of blind users. This is crucial, as it allows individuals who are blind to engage with their environment more effectively, reducing their dependence on traditional aids or human assistance.

In pursuing this research, we not only aim to enhance technological capabilities but also to champion the integration of these technologies into society, paving the way for more accessible and inclusive environments. This paper details the method-

ology, experiments, and implications of our innovative approach, setting a precedent for future studies in the field of AI-driven navigation aids.

2 Literature Review

Traditionally, 2D maps posted in common areas are used to help individuals navigate complex environments. Signs, and colour coded departments are among the methods used to help individuals distinguish areas within buildings. However, not all 2D maps are standardized, and signs and colour schemes vary building to building. Interactive maps have become more common, presenting individuals with pathways to their desired destination. Interactive maps are growing in popularity but have yet to receive widespread adoption. As well, effective use of the interactive map would require the individual to memorize the map, along with maintaining their orientation relative to the mapped directions. Maps aim to guide individuals to the proximity of the desired location, where they can then rely on signing or assistance from staff. But in order to reach the general vicinity of the desired location requires the ability to understand and utilize the provided map. Unfortunately, not all members of society, for a variety of reasons, are able to effectively use maps.

2.1 Assistive Technologies for People with Disabilities

Many groups of people within society have difficulty navigating both known and new environments. The most common afflicted group being individuals with visual impairment, but they are not alone. It is well-documented that precision in remembering spatial configurations and landmarks declines with age [8]. As well it is believed that the decline in the hippocampus of older members of society may increase their difficulty in navigating. This affects their ability to plan and impacts their choice in navigation strategies chosen, [5]. Studies have also shown that individuals who have experienced brain-injuries may develop a disorder termed “topographical disorientation”, [1]. Resulting in the selective loss in the ability to orient themselves or find their way within environments.

Technological advancements over recent years has further progressed technologies ability to empowering and assistance individuals in daily life task. Research conducted within hospital settings has highlighted how AI can be used to improve users experiences by making real-time adjustments to interactive platforms [6]. As well, the use of LLM’s have shown promise for aiding in serving rural communities by acting as a booking system [16]. With respect to navigation, assistive technologies have been developed over recent years to aid with navigating indoors spaces. Companies such as Augmented Pixels [15] and Array.ai [2] implement augmented reality for navigation. Maps can be loaded onto the platforms, while users are tracked in real time through mobile app or web services. Thus, users can use the

interface to select destinations and receive step-by-step guides. Depending on the technology, corporations typically have the options to track users in house, or via the platform. QR codes inside the building can be used to anchor users' positions with facilities. Though growing in popularity, the technology is still relatively new and has yet to receive major implementation.

2.2 Role of Artificial Intelligence in Navigation

Advancements in AI technologies and LLM's have resulted in the refinement of Vision-Language Navigation. LLM's have enabled the progression of models utilizing pre-trained vision-language methods to now being empowered by large scale data sets [9]. GPT has been used to aid in navigation, as seen by the research team behind NavGPT. NavGPT uses the BLIP-2 model [19] to translate images into text, which they then pass to GPT in order to accomplish navigation tasks. The Discuss-Nav team utilized MLM InstructBlip to act as the "scene observation expert" while using GPT-4 to act as the "decision testing expert", [11]. GPT-4 was also prompted to complete tasks such as "evaluate the feasibility of each movement prediction based on thought process and current environment".

Recently a research team launched a model called MapGPT, the model was provided with a top down map view consisting of nodes. The nodes were used to illustrate the paths for navigating between areas to GPT-4. "Specifically, [they] build an online map and incorporate it into the prompts that include node information and topological relationships, to help GPT understand the spatial environment" [4]. The goal of their work was "a novel map-guided prompting method, which introduces an online linguistic-formed map including node information and topological relationships to encourage GPT's global exploration".

2.3 Gaps and Limitations in Current Research

Indoor navigation is a popular topic of discussion, with new research and technologies emerging to aid people in navigation. An example of such technology is Augmented Pixels. The application takes advantage of the popularity of mobile devices and emerging augmented reality technology to create 3D interactive maps. Research is also being conducted to utilize the growth and development of AI technology. These research experiments implement AI using trained models or topological maps to enhance the ability of AI models to carry out navigational tasks. As previously stated, MapGPT uses GTP-4 loaded with topological data to navigate paths. Most of these systems focus on charting paths from the user to the desired locations. However, these methods do not address situations where prior knowledge is not available. Not all organizations wish to have internal networks for navigation, as internal positioning systems can present security risks.

Additionally, many organizations may prefer not to use third-party corporations to host their navigational network. Addressing these concerns would require the need for a stand-alone system, technology that requires no data from the buildings that are to be navigated. The previous methods do not address the need for an untrained AI model, with no topological information. A system that does not require data increases the possibilities in which AI becomes a viable solution for indoor navigation. Thus, making it possible for individuals to use AI for any scenario they need, as opposed to only buildings available within applications or websites.

3 Methodology

3.1 Overview of the LLM Technology

GPT4-V is an advanced iteration of OpenAI’s generative pre-trained transformer models, specifically enhanced to include vision capabilities. Research using GPT4-V for face detection in hospital settings have proven not only the effectiveness of the model, but also illustrating the opportunities for AI in medical environments [13]. This hybrid model amalgamates the robust language understanding and generation abilities of GPT-4 with a sophisticated visual processing component. The integration allows GPT4-V to receive and analyze visual inputs alongside textual data, enabling it to interpret complex scenes and provide contextually relevant responses in real-time. The primary capability of GPT4-V that is central to this study is its ability to process complex visual information and translate it into descriptive, actionable language. This feature is particularly beneficial for navigation in intricate and dynamic environments such as hospitals. By understanding both static images and dynamic visual scenes, GPT4-V can describe physical layouts, identify objects and obstacles, and provide updates about environmental changes. These abilities are crucial for assisting individuals with disabilities in navigating indoor spaces safely and effectively. In the context of this research, GPT4-V’s dual processing of visual and textual data presents a significant advantage. It allows the system to perform tasks that require a detailed understanding of the environment, such as locating exits, identifying restrooms, and providing directions based on visual landmarks. These tasks are accomplished through the model’s advanced algorithms, which integrate data from multiple sources to generate accurate and useful navigational prompts. This capability demonstrates a leap forward in making complex indoor environments more accessible through technology.

3.2 Prompting Strategy

The prompting strategy, “SDC”, developed for GPT4-V in the context of navigation assistance is designed to ensure comprehensive and structured documentation of navigation sessions. This strategic approach is crucial for creating an effective feedback loop and for refining the AI’s performance over time. Here’s how the strategy is structured:

SDC: To facilitate thorough documentation and analysis of navigation decisions, the system uses structured prompts to capture every key aspect of a navigation session. These prompts guide the AI in systematically recording decision points, paths taken, branches (both explored and not), and visual clues. This structured approach ensures that all relevant data is captured in a way that is easy to analyze and reference, based on the features below.

- **Decision Points:** The AI is prompted to identify and record decision points within the environment. Each decision point is given a unique identifier and is described in detail. This helps in understanding the choices available at each juncture of the navigation path.
- **Paths Taken:** For every path selected, the AI records detailed information including identifiers for decision points and branches, reasons for path selection, and precise timestamps. This data is crucial for tracking the sequence and timing of navigation decisions and for understanding the navigational logic of the system.
- **Branch Recording:** The AI records all potential navigation branches at each decision point, documenting whether each branch was selected for navigation. This includes a description of each branch to provide context on the navigation environment.
- **Visual Clues Documentation:** Corresponding to each branch, visual clues are documented with detailed descriptions and their potential utility in navigation. This helps in correlating visual elements with navigation decisions, enhancing the AI’s ability to utilize visual information effectively.

The tables 1-4 represent the tables used for the SDC strategy. Sample data for the tables will be based on the following scenario: You are tasked with navigating through a hospital to reach the radiology department. You are starting at the entrance of the main corridor.

In this scenario, as the user approaches each decision point, GPT4-V evaluates the options based on current user location, visible signs, and foot traffic conditions. The AI selects the most suitable paths and records each decision systematically in the tables, providing clear and useful navigational prompts based on both pre-defined criteria and real-time environmental analysis.

Table 1: Decision Points Table

Decision Point ID	Location Description
DP1	Entrance to the main corridor
DP2	Intersection near the cafeteria

Table 2: Paths Taken Table

Step ID	Linked Decision Point ID	Linked Branch ID	Reasoning	Timestamp
S1	DP1	B1	Least crowded route chosen	2023-10-01 10:00:00
S2	DP2	B2	Signage clearly visible	2023-10-01 10:05:00

Table 3: Branches Table

Branch ID	Linked Decision Point ID	Branch Description	Selected (Y/N)
B1	DP1	Turn left to radiology	Y
B2	DP2	Straight to cafeteria	Y

Table 4: Visual Clues Table

Clue ID	Linked Branch ID	Description	Navigation Use
C1	B1	Sign pointing left	Directs towards radiology
C2	B2	Large cafeteria sign visible	Helps identify destination

3.3 Role of AI in Navigation Assistance

GPT4-V significantly enhances the ability of individuals with disabilities to navigate complex indoor environments such as hospitals. The system leverages its advanced vision and language processing capabilities to provide detailed, contextual navigation assistance tailored to the specific needs of users with visual and mobility impairments.

Assistance for Individuals with Visual Impairments: For visually impaired users, GPT4-V transforms visual information into detailed verbal instructions, en-

hancing their perception of surroundings through audio descriptions. The system can articulate the layout of spaces, pinpointing locations of doors, restrooms, and other key landmarks. It also provides descriptions of obstacles and their exact locations, helping users avoid them. Moreover, in crowded settings, GPT4-V describes the density and movement of foot traffic, guiding users through less congested paths which improves their navigation experience and safety.

Support for Individuals with Mobility Disabilities: Users with mobility disabilities benefit from GPT4-V's ability to suggest accessible routes that accommodate their specific needs. The system actively identifies and recommends pathways that avoid physical barriers such as stairs and narrow corridors. It incorporates real-time updates about the operational status of essential facilities like elevators and automated doors, directing users to available alternatives whenever necessary. Additionally, GPT4-V adapts its guidance based on real-time environmental changes. For example, if an accessible route is temporarily obstructed, the system promptly recalculates and provides a new route, ensuring continuous accessibility. Through these integrations, GPT4-V not only facilitates easier and safer navigation for individuals with disabilities but also fosters a sense of independence by enabling them to move more freely within complex spaces. The role of AI in this context is not merely functional but transformative, offering a bridge between technological innovation and enhanced quality of life for users with diverse needs. This underscores the critical importance of developing inclusive technologies that cater to a wide range of abilities, ensuring that advancements in AI directly contribute to societal benefits.

3.4 Privacy and Security Concerns

In developing a vision-based Large Language Model framework for navigation in environments like hospitals, integrating privacy and security from the outset is critical. The system architecture handles this through the following aspects:

1. The framework leverages edge computing to process the bulk of visual data directly on the devices used for navigation assistance. Tasks such as initial image capture, primary object recognition, and the derivation of immediate navigational commands are handled locally. This local processing not only minimizes latency, enhancing user experience, but also strengthens data security by substantially reducing the amount of sensitive data transmitted over external networks.
2. Any sensitive information extracted from images, particularly identifiers that could be traced back to individuals, undergoes an anonymization process, like k-anonymity or differential privacy [14], before it is stored locally or incorporated into feedback data sent to servers.
3. User feedback is essential for the iterative improvement of the navigation system. Feedback data sent to central servers is primarily focused on user experiences and system performance metrics, rather than specific visual data.

4. To ensure the security of data, both locally stored and transmitted data are protected using Advanced Encryption Standard (AES) [12] with a 256-bit key. Additionally, during transmission, the data is secured further using Transport Layer Security (TLS) [12] protocols.

This integrated security framework ensures that our system adheres to the highest standards of data security, crucial for its application in sensitive environments like hospitals.

4 Experiment

The purpose of the experiments was to compare the effectiveness of three different prompting strategies in a vision-based LLM navigation system within a hospital environment. These three strategies were tested against challenges of varying degrees of difficulty. The first challenge was a simple instruction: turn right, travel to the end of the hall, and enter the room on your right. The next challenge added difficulty by instead of providing a direction, only providing a destination, but informing GPT4-V that the destination was on the same floor. The remaining challenges did not indicate which floor the destination was on, increasing the difficulty of the challenge. All of these tests were performed to test the effectiveness of the different strategies for executing the desired outcome.

For each experiment, GPT4-V was instructed with its current location, the destination, the chosen strategy, and a series of images from a first-person perspective inside the hospital. These strategies are meant to illustrate GPT4-V prowess in analyzing images and applying reasoning in the decision-making process. While demonstrating GPT4-V limitations in understanding and navigating new complex 3D environments. The three implemented strategies for navigating the hospital are as follows:

Baseline: Provide GPT4-V with a set of images representing the user perspective within the physical environment, (generally Front, Left, and Right where possible). Then prompt GPT4-V for which direction to travel based upon reaching the instructed destination. Finally traverse the space until a new decision point is reached, (decision point: an area with either multiple directions to travel such as an intersection, or a dead end such as a room at the end of the hall). For example, following GPT4-V’s decision at a junction on the path to travel next, the experiment would continue on the forward path until a new decision point is reached. This iterative process of prompting GPT4-V at each decision point would be continued until either the final destination is reached, or the user finds themselves stuck in a loop, continuously traveling between previously visited decision points.

Image analysis: Similar to the Baseline strategy, the user would prompt GPT4-V with a set of images based on the current perspective, querying for the direction to travel next. The difference is that before making a decision as to which direction to travel to next, a description of the analyzed image is provided by GPT4-V. This description would include any landmarks, identifiable features, and other specifics

that would aid in navigation (landmarks, identifiable features, and specifics would be decided by GPT4-V itself). The purpose of having GPT4-V provide breakdowns of the images was to use them as references for future decisions. GPT4-V is stateless, meaning that it does not retain information and reads the conversations from the most recent prompt to the beginning or maximum token allotment [10]. This means old images from previous decision points are not being re-analyzed, so crucial visual cues from previous decision points were missed using the Baseline strategy. By creating text copies of visual cues inside the conversation, the Image Analysis strategy attempts to maximize GPT4-V’s visual analysis ability by ensuring that potentially vital information is carried forward for future decisions.

SDC: The final strategy placed a heavy emphasis on tracking the paths encountered by GPT4-V during the experiment. A set of 4 tables was used to track the path, decisions, and observations of the experiment. Specifically, a decision points table is needed to track all the encountered decision points. A branch table is used to track all the possible directions based on the available paths leading out from decision points. A path table is used to track the path followed during the course of the experiment. Finally, a visual cue table: to make note of key features associated with each branch. Upon reaching a decision point, a set of images would be uploaded. GPT4-V analyzes the images, then makes a decision on the next course of action. Finally, GPT4-V would update all 4 tables to reflect the latest decision in navigating the environment. The Image Analysis strategy helped GPT4-V make better decisions in comparison to the Baseline strategy. However, at dead ends, GPT4-V using the Image Analysis strategy struggled to accurately return to previous decision points or determine already explored paths. The tables in the SDC strategy provided a clearer format for GPT4-V to analyze the current path when deciding which decision points to recursively return to on the table.

4.1 Experimental Design

The Nile of Hope Hospital [3], a seven floor hospital in Alexandria, Egypt was the experiment site. The Nile of Hope Hospital was selected due to its first person virtual environment via Matterport 3D. The lack of commodities such as restaurants or stores removed distractions and possible noise from the collected data. Finally, the compact design of relatively small stacked floors, instead of long wings aided in the rapid testing of the experiment protocols. The compact floors resulted in smaller areas and therefore smaller gaps between decision points. The general layouts of the floors were open concept, with colourful waiting areas, and rooms along the perimeter. The colourful waiting areas were particularly important as they were a distinctive landmark that helped to anchor the relative position of GPT4-V within the hospital. An additional noteworthy feature was that the hospital did not have lines along the floor or walls to indicate paths to specific departments. This ensured that GPT4-V had to rely on visual cues and reasoning to determine the direction to travel to reach the desired destination. The experiment consisted of four testing con-

ditions for each strategy. Testing GPT4-V's ability to understand and navigate the environment, assess its reason and ability to reach the desired destination. Illustrating the abilities and limits of GPT4-V for navigation in new complex environments. The test cases are as follows:

1. The experiment starts on the top of the staircase entering the 3rd floor. The goal of the experiment was to navigate to turn right, travel to the end of the hall, and enter the room directly to the right of the current perspective. The aim of the prompt was to demonstrate GPT4-V's ability to correctly analyze the images and gauge GPT4-V's ability to use judgement.
2. Instead of providing GPT4-V with directions, only a destination— radiology department— was provided. GPT4-V was only tasked with locating the radiology department, not a specified room or type of equipment. The task was deemed as a success, if GPT4-V navigated to the radiology department and correctly identified that it had indeed located the desired destination. The starting location is the same as task 1. The third floor is also the location where the radiology department resides inside the hospital. GPT4-V was informed in the initial instructions that the destination was located on the same floor.
3. Similar to test case number two, the desired destination was the radiology department on the third floor of the hospital. However, the starting location was the staircase on the entrance to the second floor instead of the third floor. GPT4-V was then apprised of its starting location, and informed that the radiology department was on either the second or third floor. The experiment was deemed a success when GPT4-V had navigated to the desired destination and correctly identified having reached the radiology department. Only staircases that had been encountered on the current or previous decision points could be used for traveling between floors. This was to reflect real-life navigation, as individuals can only use staircases which they have direct access.
4. The test case was the same as task 3, the only difference being that the floor level of the department resided was not provided. Instead, the initial instructions only provided the information that radiology was not on the first floor. Successful completion of the experiment would include the navigation and correct identification of the radiology department once reached. As with test case three, only encountered staircases could be used to travel between floors.

The setup of the experiment is an initial prompt to GPT4-V indicating the experiment goals, strategy and the first set of images representing the starting location. The specifics of the text prompt will vary depending on the experiment. The following is the text prompt for task number one implementing the Baseline strategy: "You are tasked with navigating to the right, and traveling to the end of the hall. Once you have reached the end of the hall, enter the room to the right. To assist you, three images have been uploaded, representing views [Forward], [Left], and [Right] from your current position. Based on these perspectives, decide which direction to proceed". The criteria used for measuring performance is the successful completion of the specified task. The task will vary depending on the test, but each task includes arriving, navigating to the desired destination, and the ability to correctly identify

having arrived at the destination. Outside the main measurable, observations were taken to compare:

- The decisions made and the reasoning behind decisions.
- The ability to back track between decision points.

The measured performance aims to demonstrate the effectiveness of the SDC strategy in comparison to the Baseline and Image analysis Strategies. While the observations provide an in-depth look at the underlying decision-making process for developments and future research.

4.2 Results

For the experiment, each task was 4 separate times for each strategy (Total 48 tests). Table V illustrates the results of the test, for each strategy two data point were measured. The first is the number of the percentage of successful completions of each task based on the previously defined criteria. The second value is the average number of prompts per test case. A prompt for the test was defined as each instance information was sent to GPT4-V. Regardless of size, or whether the data contained text, images or both. The significance of measuring the number of prompts is to illustrate the increase of level of difficulties between task and compare each strategy's ability to maintain cohesion and awareness of the space as the complexity increased between task. Higher prompts number meant a process was more complex, but also that the strategy was capable of maintaining its awareness within the environment.

Table 5: Experiment Comparison

Task	Baseline		Image Analysis		SDC	
Task 1	75%	5	75%	5.25	75%	4.75
Task 2	0%	4.5	0%	5.25	100	9.75
Task 3	25%	7.5	0%	7.5	25%	8.25
Task 4	25%	4.5	25%	5.5	50%	7.67

GPT4-V image analysis ability proved to be relatively accurate regardless of the selected strategy. This is illustrated by the first task of traveling to the end of the hall, as each strategy was able to navigate the area and clearly state when the destination was reached.

For the second test, GPT4-V started on the third floor, with the goal of arriving at the radiology department (the radiology department is also located on the third floor). For task two the Baseline strategy proved ineffective, having a completion rate of zero. For each test conducted, GPT4-V would select a wrong path, and without the ability to accurately back track each navigation was unsuccessful. The

experiment would generally end either by GPT4-V suggesting to travel beyond dead ends or not being able to identify previous decision points to back track.

The Image Analysis strategy improved upon the Baseline strategy in practice, as noted by the increase in the average number of prompts. But the end result was the same, resulting in a completion rate of 0%. The notable difference between the two initial strategies were the ability of GPT4-V to accurately backtrack to previous decision points. The Image Analysis strategy showed promise as during one test GPT4-V was able to complete the task after twice selecting the wrong path. After following both paths to dead ends, the Image Analysis strategy returned to previous decision points. But the complexity of the environment required the ability to backtrack numerous decision points. This proved to be too difficult, thus proving a minor increase in effectiveness over the Baseline but an insufficient end result.

The SDC strategy proved to be a significant leap in efficiency. Having a 100% completion rate in task 2. But the jump in success rate came at the cost of efficient, an average number of 9.75. Double the amount of the Baseline Strategy and almost double the Image Analysis strategy. It should also be noted that while the proposed strategy has proven to be effective in backtracking task two, this effectiveness is lost in more complex areas such as those in task three. While the third floor had an average back track of 2 decision points between dead ends and starting points, task three had upwards of 6 for certain paths. The increase in complexity can also be seen in the increase in the average number of prompts carried out by each task. This is the assumed reason why the SDC strategy rate reduction, as backtracking would in task three eventually lead to GPT4-V becoming caught in loops of repeated traveling between already discovered decision points. While the SDC strategy has shown promise in creating a framework for future navigation in complex environments, its performance in tasks 3 and 4 highlighted some challenges. The increased complexity of these tasks led to significant reductions in the strategy's effectiveness, with success rates dropping to 50% and 25%, respectively. This was primarily due to the tables growing in magnitude and complexity, which sometimes caused the system to loop repeatedly between already discovered decision points.

4.3 Discussion

The Image Analysis and the SDC strategy were created to improve GPT4-V navigational abilities, but it could be argued that the Image Analysis and the SDC strategy are essentially simulating memory. GPT4-V simulates memory within conversations by reading the entirety of conversations before responding. As previously noted the ability to review the conversation only applies to text. Images are analyzed in real-time, meaning only uploaded images reflecting the current perspective are analyzed with each prompt. The Image Analysis strategy takes a step toward human-like memory by making note of distinct features within the environment. In the same way an individual would take note of notable features within new environments, GPT4-V captures these notes within the conversation.

To the left, there's a large purple and blue sign with Arabic writing, possibly providing directions or department names. Straight ahead, there is a seating area with yellow and green chairs, and a decorative wall with a mountain and balloon motif, likely designed to create a comforting atmosphere for children. On the right, there is a series of doors with windows, one marked with a warning for an electrical room (UPS). A red fire hose reel is mounted on the wall next to the door. Overhead, there are green exit signs indicating the way out.

This is an excerpt taken from a response of one of the experiments utilizing the Image Analysis strategy. Throughout the response, colours and common objects such as signs, windows and doors are noted. When combined, the summation of the notable features allow for the creation of mental images to aid in guidance and recognition of distinct areas. The inclusion of image descriptions dramatically increased the navigation success rate, as seen in the improvements between the Baseline and the Image Analysis strategy. In the same way, an individual might create mental images, associating objects or colours with different areas. The Image Analysis strategy mimics the act of mental images by noting each pathway with a series of notable features for future reference.

As noted in the results section, the Image Analysis strategy's effectiveness decreased as the complexity of the navigating environment increased. This is, in part, due to a distinctive difference between the Image Analysis strategy and how people navigate environments. Visual cues, mental images and navigation's algorithms are capable of allowing individuals to navigate closed environments. However, in order to effectively navigate new complex environments, individuals need to create mental maps. Mental maps are an individuals unique perception and spacial understanding of their environment [18]. The SDC strategy combats this by creating tables to track the path taken, noting district features and tracking all encountered branches. By having the updated path tracking tables within the conversation, for each prompt GPT4-V is provided with the information to create a map of the environment. Effectively giving GPT4-V a mental map to use for making decisions. While at first seeming more mechanical, the ability to understand the path navigated, how branches are connected, and associating areas with notable features provide GPT4-V with a sense of visual memory and a mental map closer to humans.

An added benefit of the tables is the ability to dynamically update the tables as needed to correct information. In the same fashion that individuals can correct inaccuracies in their mental maps, GPT4-V dynamically updates its tables to address any errors in the navigation data. For instance, if GPT4-V incorrectly identifies an element in an image, it can update the tables to reflect the correct information in future interactions. This capability is vital as it prevents navigation errors that could arise from outdated or conflicting data, ensuring that the most current and accurate information is always used for decision-making. This process closely mimics human memory, where people continually update their understanding of their surroundings based on new information.

As GPT continues to progress, the hope is by using a refined version of the SDC strategy to aid individuals in navigating new environments in real time. Potentially for individuals with visual impairments, by using the SDC strategy any individual should be able to navigate hospitals, airports, and more all from their cellular device.

Before practical applications, the problem of table maintenance would need to be addressed. As the complexity of the navigation task increases, the complexity and length of the table proportional increases. Rather than having to rebuild the table following each decision, a solution would be to utilize a separate database where data could be stored, retrieved and updated. With advancements in AI and further refinement of the navigational process, this research displays the potential GPT has in aiding individuals in navigating their everyday lives.

5 Conclusion

This study has systematically evaluated the efficacy of a novel prompting strategy, SDC, in a vision-based LLM for indoor navigation, comparing it against two other methods: one that simply provides pictures for navigation, and another that provides pictures along with environmental descriptions. The experiments conducted in a hospital setting with tasks of increasing complexity have demonstrated that our novel strategy not only consistently outperforms the other methods but also completes all designated tasks with notable efficiency and accuracy. The key findings from our experiments highlight the superior capability of the novel prompting strategy to utilize dynamic environmental information effectively, resembling human-like navigation more closely than traditional tree-based search methods. This approach has shown a remarkable ability to adapt to changes and correct navigational errors in real-time, leveraging visual and contextual cues to enhance the decision-making process. The impact of this strategy on enhancing navigational aids is particularly significant for visually impaired individuals, as it offers a more intuitive and accessible way to navigate complex environments like hospitals. The strategy's success in providing detailed, contextual, and adaptable navigational prompts could greatly increase the autonomy and confidence of users with visual or mobility impairments, potentially transforming their interaction with such environments.

Based on the experimental findings, there are several avenues for further enhancing the novel prompting strategy. For instance, incorporating more granular feedback mechanisms to fine-tune the AI's decision-making process could further improve its performance. Additionally, integrating multi-modal feedback from users, such as voice commands or gesture recognition, might provide richer interaction data for the AI to learn from. Future research could also explore the application of this prompting strategy in other complex environments, such as airports, shopping malls, or university campuses, where navigation can be equally challenging. Examining the effectiveness of the strategy across different user demographics, including varying levels of disability and technological proficiency, would provide deeper insights into its adaptability and user-friendliness. Moreover, the scalability of this novel strategy suggests its potential applicability to other types of AI-driven assistance systems, such as autonomous vehicles or robotic aides, where dynamic environment interaction is crucial. The principles established through this research

could inform broader AI applications, making a significant impact on the development of inclusive technology solutions.

Acknowledgements

The authors wish to express their sincere gratitude to Algoma University Research Fund for their generous support of this work. Their commitment to fostering innovative research has been invaluable to the success of this project.

References

1. Aguirre, G.K., D'Esposito, M.: Topographical disorientation: a synthesis and taxonomy. *Brain* **122**(9), 1613–1628 (1999)
2. ARway.ai: Arway.ai introducing AR navigation with generative AI GPT 4D avatars in new partner deal with avr labs in UAE (2024). URL https://www.arway.ai/press_releases/arway-ai-introducing-ar-navigation-with-generative-ai-chatgpt-4d-avatars-in-new-partner-deal-with-avr-labs-in-uae/. Accessed: 2024-05-05
3. Chang, A., Dai, A., Funkhouser, T., Halber, M., Niessner, M., Savva, M., Song, S., Zeng, A., Zhang, Y.: Matterport3D: Learning from RGB-D data in indoor environments. *International Conference on 3D Vision (3DV)* (2017)
4. Chen, J., Lin, B., Xu, R., Chai, Z., Liang, X., Wong, K.Y.K.: Mapgpt: Map-guided prompting with adaptive path planning for vision-and-language navigation. *arXiv preprint arXiv:2401.07314* (2024). URL <https://doi.org/10.48550/arXiv.2401.07314>. Submitted on 14 Jan 2024, last revised 25 Feb 2024
5. Cogné, M., Taillade, M., N'Kaoua, B., Tarruella, A., Klinger, E., Larrue, F., Sauzéon, H., Joseph, P.A., Sorita, E.: The contribution of virtual reality to the diagnosis of spatial navigation disorders and to the study of the role of navigational aids: A systematic literature review. *Annals of Physical and Rehabilitation Medicine* **60**(3), 164–176 (2017)
6. Ghosh, A., Huang, B., Yan, Y., Lin, W.: Enhancing healthcare user interfaces through large language models within the adaptive user interface framework. In: *Information and Communication Technology: ICICT 2024*, pp. 1–10. Springer (2024)
7. Ghosh, A., Yan, Y., Lin, W.: Adaptive user interface framework powered by a large language model for culturally sensitive virtual healthcare applications. In: *IEEE International Conference on Biomedical and Health Informatics (BHI'23)*. Pittsburgh, Pennsylvania, United States of America (2023)
8. Head, D.: Age effects on wayfinding and route learning skills. *Behavioural Brain Research* **209**(1), 49–58 (2010)
9. Li, F., Zhang, H., Zhang, Y.F., Liu, S., Guo, J., Ni, L.M., Zhang, P., Zhang, L.: Vision-language intelligence: Tasks, representation learning, and large models. *arXiv preprint arXiv:2203.01922* (2022). DOI 10.48550/arXiv.2203.01922. URL <https://arxiv.org/abs/2203.01922>. Submitted on 3 Mar 2022
10. Lingard, L.: Writing with chatgpt: An illustration of its capacity, limitations & implications for academic writers. *Perspectives on Medical Education* **12**, 261–270 (2023)
11. Long, Y., Li, X., Cai, W., Dong, H.: Discuss before moving: Visual language navigation via multi-expert discussions. *arXiv preprint arXiv:2309.11382* (2023). URL <https://doi.org/10.48550/arXiv.2309.11382>. Submitted to ICRA 2024

12. OpenAI: Enterprise privacy at openai. URL <https://openai.com/enterprise-privacy/>. Accessed: 2024-05-05
13. Rosario, I., Huang, B., Yan, Y., Lin, W.: Enhancing telehealth patient experience with emotion-sensitive large language models. In: Information and Communication Technology: ICICT 2024, pp. 1–10. Springer (2024)
14. Sebastian, G.: Privacy and data protection in chatgpt and other ai chatbots: Strategies for securing user information. International Journal of Security and Privacy in Pervasive Computing (IJSPPC) **15**(1), 1–14 (2023)
15. The Odessa Journal: Icu invest in augmented pixels (2020). URL <https://odessa-journal.com/icu-invest-in-augmented-pixels>. Accessed: 2024-05-05
16. Xia, M., Huang, B., Yan, Y., Lin, W.: Transforming patient experience in underserved areas with innovative voice-based healthcare solutions. In: Information and Communication Technology: ICICT 2024, pp. 1–10. Springer (2024)
17. Zhang, T., Hu, X., Xiao, J., Zhang, G.: A survey of visual navigation: From geometry to embodied AI. Engineering Applications of Artificial Intelligence **114**, 105,036 (2022)
18. Zhang, W.: Mental map: A reliable definition of choice or a distorted recognition of space? Degree project, Transport and Location Analysis, Stockholm, Sweden (2015). PDF, 62 pages, June 17, 2015
19. Zhou, G., Hong, Y., Wu, Q.: Navgpt: Explicit reasoning in vision-and-language navigation with large language models (2024)