

“© 2026 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works.”

Guide-LLM: An Embodied LLM Agent and Text-Based Topological Map for Robotic Guidance of People with Visual Impairments

Sangmim Song¹, Sarath Kodagoda¹, Amal Gunatilake¹, Marc G. Carmichael¹
Karthick Thiyagarajan² and Jodi Martin³

Abstract—Navigation presents a significant challenge for persons with visual impairments (PVI). While traditional aids such as white canes and guide dogs are invaluable, they fall short in delivering detailed spatial information and precise guidance to desired locations. Recent developments in large language models (LLMs) and vision-language models (VLMs) offer new avenues for enhancing assistive navigation. In this paper, we introduce Guide-LLM, an embodied LLM-based agent designed to assist PVI in navigating large indoor environments. Our approach features a novel text-based topological map that enables the LLM to plan global paths using a simplified environmental representation, focusing on straight paths and right-angle turns to facilitate navigation. Additionally, we utilize the LLM’s commonsense reasoning for hazard detection and personalized path planning based on user preferences. Simulated experiments demonstrate the system’s efficacy in navigating indoor environments based on user queries, underscoring its potential as a significant advancement in assistive technology. The results highlight Guide-LLM’s ability to offer efficient, adaptive, and personalized navigation assistance, pointing to promising advancements in this field.

Code and replication details: <https://github.com/awd1779/Guide-LLM>

Index Terms—Assistive Robotics, Large Language models, Text-Based Topological Mapping, Vision and Language Navigation, Safety in HRI, Vision Impairments.

I. INTRODUCTION

Navigating everyday environments can be particularly challenging for persons with visual impairments (PVI), who often depend on specialized tools, help from others, or familiar routes to get around [1], [2], [3], [4], [5]. Traditional aids, such as white canes and guide dogs, are essential parts of the navigation; however, with the technological developments, further assistance could be possible to improve user confidence in navigation [6].

Recent breakthroughs in artificial intelligence, especially in large language models (LLMs) [7], [8] and vision-language models (VLMs) [9], [10], [11], have created new

This research was supported by the Australian Government through the Australian Research Council’s Linkage Projects funding scheme (LP220100430) and the industry partner Guide Dogs NSW/ACT.

¹Sangmim Song, Sarath Kodagoda, Amal Gunatilake and Marc G. Carmichael are with the Robotics Institute, Faculty of Engineering and Information Technology, University of Technology Sydney, Broadway, Ultimo NSW 2007, Australia. Email: Sangmim.Song@student.uts.edu.au

²Karthick Thiyagarajan is with the Smart Sensing and Robotics Laboratory (SensR Lab), Centre for Advanced Manufacturing Technology (CfAMT), School of Engineering, Design and Built Environment (SoEDBE), Kingswood, NSW 2747, Australia.

³Jodi Martin is with Guide Dogs New South Wales / Australian Capital Territory, Sydney, New South Wales, Australia.

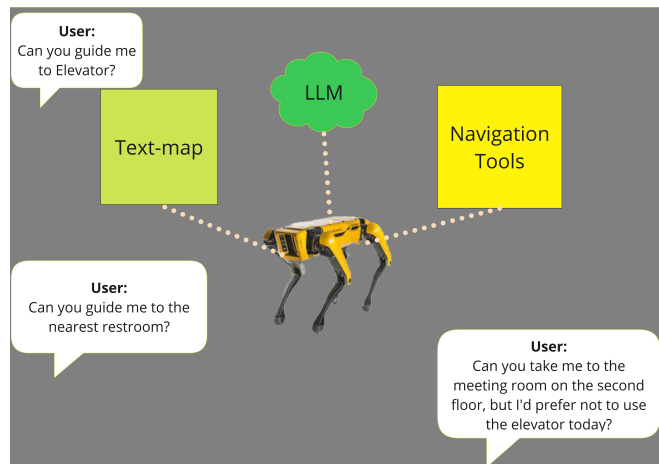


Fig. 1. Guide-LLM: The embodied agent consists of a text map, LLM, and navigational modules to guide the user to the destination.

opportunities in human-robot interaction [12], task planning [13], [14], and navigation [15], [16]. Despite these advancements, their application in assisting PVI with navigation is still largely unexplored. [17] demonstrated the potential of using a robot platform combined with a language model to assist PVI, offering a glimpse of what is achievable. However, fully harnessing the capabilities of LLMs and VLMs for guiding PVI remains largely unexplored.

Conventional navigation systems typically depend on pre-programmed rules and sensor data, which may overlook the subtleties and complexities of real-world environments. In contrast, LLMs can analyze contextual information and anticipate potential hazards, offering a more adaptive and responsive solution for navigation. A key challenge in creating LLM and VLM-based navigation systems for PVI is their dependence on precise, explicit commands from users, which can be difficult for PVI. While 3D reconstruction techniques using point clouds [18], [19], [20] and traditional methods like SLAM (Simultaneous Localization and Mapping) help LLMs understand the environment, their scalability is constrained by the high computational demands of having LLMs interpret these dense maps [21], [22], [20].

To overcome this challenge, we propose an innovative framework that utilizes a text-based topological map of the environment. This allows the LLM to plan global paths by referring to a textual representation, eliminating the need for explicit user input. This approach is more computationally efficient and scalable compared to methods that rely on

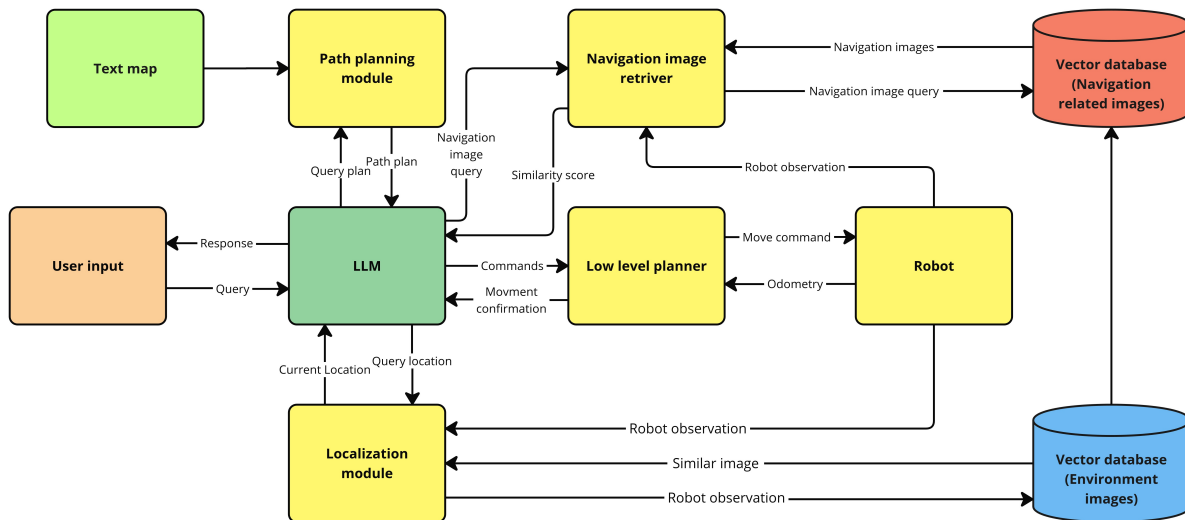


Fig. 2. Guide-LLM framework: LLM (green) serves as the central controller, using commonsense reasoning to interpret user queries and interact with various modules (yellow) for decision-making and navigation tasks. Text map (Green) provides a textual representation of the environment used by the path planning module to create route plans. Vector database 1 (blue) stores static embeddings of the environment images, aiding in consistent localization. Vector database 2 (red) stores navigational image embedding that can be updated or deleted based on the agent’s requirements.

dense maps or 3D representations for each user query, which could cause delays, leaving PVI waiting for the LLM to process these complex inputs. Additionally, our text-based topological map is designed to address the specific needs of PVI by generating straight paths and right-angle turns, which are easier to navigate and help maintain spatial orientation [23], [24]. These clear and predictable routes reduce cognitive effort, enabling more efficient and safer navigation, especially when compared to the challenges posed by curved or irregular pathways.

Guide-LLM also incorporates an image retrieval system for localization and a low-level planner to handle robot movement, constraining the robot’s actions to predictable patterns. One of the significant advantages of integrating LLMs into navigation is their ability to leverage commonsense reasoning, which has the potential for enhancing safety, personalized navigation, and the interpretability of its actions [16], [25], which traditional navigation systems often lack. The main contributions of our work are:

- We propose Guide-LLM, an innovative framework that employs LLM as embodied agents to assist visually impaired individuals in navigation.
- By integrating a novel text-based topological map, LLM, and navigational tools (Fig. 1), our approach enables high-level planning while minimizing the need for users to provide extensive descriptions of destination.
- In comprehensive simulations, we validate our method’s localization error detection and recovery, personalization potential, and hazard detection demonstrating the robust capabilities required for guiding PVI.

II. RELATED WORK

A. Navigation Aid for People with Visual Impairment

Recent advancements in assistive technology have greatly enhanced the navigation and mobility of PVI. For instance,

the robotic white cane [26] employs multimodal sensing and steering assistance to aid users in navigating their surroundings. Additionally, wearable systems have been developed to improve situational awareness by delivering real-time feedback through auditory or haptic signals [27], [28], [29], [30]. [31] introduced a method that allows PVIs to jog on athletic tracks. Despite these advancements, many of these systems are pre-programmed with fixed rules and behaviors, which can limit their adaptability to dynamic scenarios. To address these limitations, dialogue-based robots have been proposed as an alternative form of assistance, using conversational interfaces to help users reach their destinations [17]. However, these systems still encounter significant challenges, particularly in natural language interaction and adaptability. Their rule-based nature also limits their ability to generalize and adjust to changing environments, which is essential for offering personalized and flexible support to PVI users.

B. Large Language Models in Robotics

The integration of LLMs into robotics has demonstrated significant potential in areas such as task planning, autonomous driving, multimodal reasoning, and navigation. Frameworks like SayPlan and ReAct enhance robotic capabilities by breaking down complex instructions into actionable sub-tasks and combining reasoning with action [13], [32]. DriveLLM [16] showcased how LLMs can improve interpretability and decision-making in autonomous driving by integrating object-level vector modalities with LLMs, thus enhancing context understanding and explainability in driving scenarios. RoboVQ [33] combined LLMs with vision inputs, enabling robots to perform complex, long-horizon tasks and demonstrating LLMs’ multimodal reasoning capabilities. [34] utilized LLMs as navigation agents, leveraging their reasoning capabilities to develop search heuristics for exploring new environments. SayNav employs LLMs to

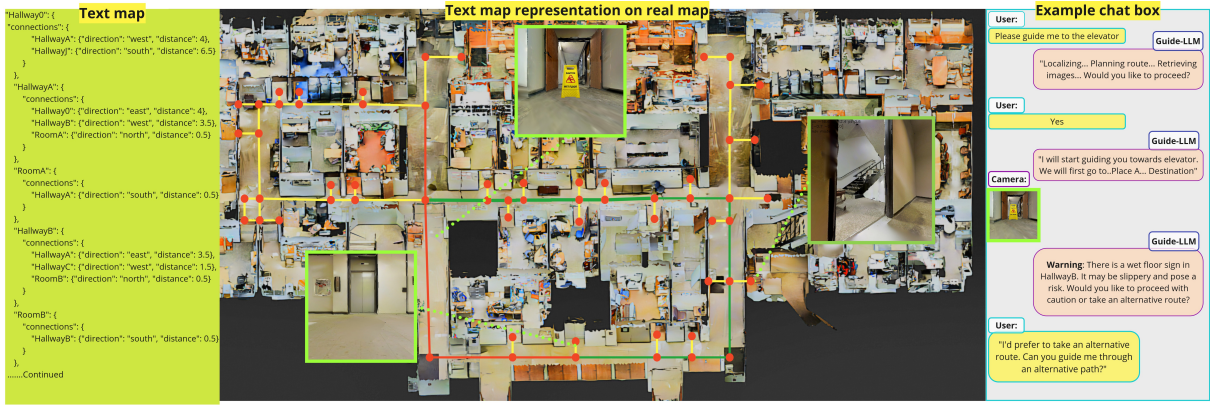


Fig. 3. Text map (Left): diagram of text map, part of the text map is extracted. Example text map representation (Middle): User asks the agent to navigate to the elevator. Guide-LLM plans a route (red line) and begins guiding. Along the route, a hazard is detected (wet floor sign), Guide-LLM warns the user and suggests an alternative path (green line), and the chat box (right) shows an example communication between Guide-LLM and the user.

guide robots through unfamiliar environments by grounding high-level instructions in spatial contexts, highlighting the potential of LLMs in exploration without requiring detailed prior knowledge [35]. [36] demonstrated LLMs’ ability to generalize across tasks, allowing robots to plan and execute complex actions with minimal task-specific training. Additionally, NavGPT [25], MAP-GPT [37], and LGX [38] explore LLM-guided robotic navigation, further extending LLM capabilities in uncharted environments. Despite these advancements, there is limited research specifically focusing on using LLMs’ commonsense reasoning and contextual understanding to assist people with visual impairments. This gap highlights the need for innovative approaches that harness LLMs’ reasoning abilities to provide more intelligent and adaptive support for PVI.

C. Vision and Language Navigation

Vision and language navigation (VLN) integrates visual perception with natural language understanding to enable agents to navigate based on spoken or written instructions. A significant challenge in VLN is that natural language instructions often emphasize high-level decisions and landmarks, frequently lacking detailed, low-level movement guidance [39]. Attention-based mechanisms [40], [41], [42] and reinforcement learning approaches [43], [44] have shown promising results in addressing this issue. Recently, the rise of LLMs and VLMs has sparked interest in leveraging these models to enhance VLN capabilities. For instance, LM-Nav [15] uses LLMs to extract landmarks from user queries and VLMs to ground these landmarks in the environment for navigation. Similarly, [45] employs LLMs to translate user queries into actionable navigation tasks and use image segmentation techniques to create a topological map for navigation. [38] combined LLMs commonsense reasoning with VLMs to navigate towards uniquely described objects, showcasing LLMs potential in understanding nuanced language. [46] utilized LLMs to guide the agent by processing action sequences based on text prompts that include navigation instructions, visual landmark descriptions, and the agent’s

past trajectory. [17] used LLMs to identify landmarks, describe the environment, and navigate through it. Despite these advancements, navigation for PVI remains challenging, particularly due to difficulties in describing specific scenes and landmarks, which can impede effective communication with the agent. Our work aims to address this limitation by enabling our agent to infer user intent and automatically resolve ambiguities in their instructions.

III. PROPOSED SYSTEM

In this section, we provide a detailed overview of the Guide-LLM framework. Figure 2 illustrates the primary components of the framework, which enable our LLM agent to handle decision-making, high-level planning, and navigation.

A. Core Components

1) *Agent*: We used GPT-4o [48] as the central component of our framework. As shown in Fig. 2, it is tasked with interpreting user inputs and determining appropriate actions to assist the user. When a user query is received, the LLM is provided with a system prompt containing instructions to process the query within a specific context. The system prompt designed to provide context to LLM which enables the LLM to perform appropriate action in each steps.

2) *Text-Based Topological Map*: To generate our text-based topological map (Fig. 3), we first construct an occupancy grid map of the area and then manually place nodes at each intersection, turn-able location, and critical navigation decision point. These nodes form the vertex set V of a directed and weighted graph $\mathcal{G} = (V, E)$, where each edge e_{ij} connects node i to node j and is annotated with a distance d_{ij} and a relative direction θ_{ij} . The straight edges that connect these nodes not only reflect direct traversable paths but also encapsulate the metric and directional information necessary for navigation. When embedded in a Euclidean plane, each node is assigned a coordinate $p_i = (x_i, y_i)$, and the spatial transformation from node i to node j is expressed by

$$p_j = p_i + (d_{ij} \cos \theta_{ij}, d_{ij} \sin \theta_{ij}).$$

TABLE I
FEATURE COMPARISON TABLE OF OUR METHOD AND OTHERS

Approach	Natural language interactions	Doesn't Require Detailed Goal Knowledge	Localization Error & Recovery	Identify potential Hazards	Personalization
Guide-LLM (Our Approach)	✓	✓	✓	✓	✓
Traditional Navigation Aid [26] [47]	✗	✗	✗	✓	✗
Wearable System [27], [28], [29], [30]	✗	✗	✗	✓	✗
VLN Approach [15] [17] [38] [45]	✓	✗	✗	✗	✗

We intentionally avoid curved paths or minor direction changes to maintain simplicity, in accordance with findings [24], [23] that suggest straight paths are more comfortable for individuals with visual impairments.

3) *Path Planning Module*: The planning module works in tandem with the agent to plan efficient routes to designated destinations. Upon a navigation query, it processes the text-based topological map, and generates a textual description of the route for the agent to process. This approach allows the agent to assess multiple paths and adapt in real-time based on hazards or user preferences.

4) *Embedding Module*: We use off-the-shelf CLIP [10] to generate embeddings from visual data, which are then stored in a vector database alongside location and orientation metadata. This metadata allows the LLM to reference the location and orientation of specific images upon retrieval. The embedding module is essential for localization and sub-goal selection, as it supports both text-to-image and image-to-image similarity searches.

5) *Vector Database*: The vector database stores high-dimensional representations of the environment generated by the embedding module, enabling real-time comparison of current observations with previously stored data. The agent accesses this data by querying navigation-related images and uses them to compare current observations with retrieved images.

6) *Low-Level Planner*: The low-level planner converts the high-level decisions made by the agent into executable actions for the robot. It ensures that these commands are physically feasible, considering the robot's movement capabilities.

7) *Robot*: We use TurtleBot as the robotic platform.

B. Global Path Planning

Figure 2 illustrates the overall process of the proposed framework. We start with the assumption that the text map and images in the main vector database are pre-labeled. The process begins with the agent processing the user's query and the system prompt. The agent then generates high-level plans, navigational image query commands, and responses for the user. As shown in Fig. 2, each output is communicated separately for clear communication between

modules. High-level plans and user responses are delivered through a voice-to-text interface, while image queries retrieve navigation-related images from the vector database. These images are then embedded into a secondary vector database, which refines the navigation process by reducing ambiguity and ensuring that only relevant images are retrieved.

C. Topological Navigation

The agent begins navigation by querying the secondary vector database to retrieve the image corresponding to the current node. The retrieved image is compared to the current observation using cosine similarity. To improve place recognition, a similarity check is performed each time the robot travels to the next node or makes a turn. The low-level planner measures the distance traveled to the next node, and upon arrival, it sends a message to the agent. If the similarity score exceeds a predefined threshold, the agent concludes that the target node has been reached and generates the next set of movement commands. This process continues until the final destination is reached.

D. Localization and Error Handling

Localization in the Guide-LLM framework employs a dual-layered approach. Initially, the agent verifies its arrival at the desired node by querying the navigational vector database and comparing the retrieved image with the robot's current observation. If the similarity score between these images falls below a set threshold, the agent detects a potential localization error, suggesting that the robot may be in the wrong location. To address this, the agent initiates a broader search by querying the main vector database, which contains embeddings of the entire environment. This fallback mechanism allows Guide-LLM to re-localize by finding the most relevant match among all environment images. This dynamic error-handling process ensures that even if an agent is initially misaligned or disoriented, Guide-LLM can correct its location in real-time, maintaining the desired path for navigation.

E. Utilizing LLM's Commonsense and Reasoning

Our framework utilizes the LLM's commonsense reasoning to improve navigation safety and decision-making.

TABLE II

LLM PERFORMANCE COMPARISON ON NAVIGATION (SUCCESS RATES)

Model (Parameters)	House	Office
GPT-4o [48]	84%	86%
Claude 3.5 Sonnet [49]	74%	78%
Llama 3.2 Vision 90B [50]	76%	74%
Llama 3.2 Vision 11B [50]	20%	24%
Qwen2.5-VL-7B [51]	0%	4%

Unlike traditional systems that depend on predefined rules, the agent can interpret dynamic, real-world contexts to anticipate potential risks. For instance, if the agent identifies hazards such as a wet floor, warning tape, or unexpected obstacles through visual data or environmental descriptions, it proactively alerts the user and suggests an alternative route. The LLM’s reasoning capabilities allow it to detect potential hazards even if they are not explicitly mentioned what is hazard. This flexibility enables the agent to adapt to changing conditions that rule-based systems might fail. By integrating commonsense knowledge, the agent enhances both the safety and overall reliability of the navigation experience.

F. Personalization Potential

A key strength of Guide-LLM is its ability to personalize the navigation experience according to each user’s specific preferences and needs. The agent’s natural language interaction capabilities enable it to adjust its behavior in real-time, facilitating this personalization. For example, it can modify its path-planning to match a user’s preferred walking speed, route preferences (e.g., avoiding stairs or choosing quieter areas), or specific safety concerns, such as steering clear of potential hazards. Additionally, Guide-LLM can incorporate feedback from previous interactions, progressively refining its decisions to align with the user’s habits and preferences. For instance, if a user consistently opts for longer, less crowded routes over shorter ones, the agent can integrate this preference into its future planning. This adaptability extends beyond navigation; the agent can engage in personalized dialogue, adjusting the level of detail and communication style to meet different users’ needs, whether they prefer brief instructions or more detailed explanations.

IV. RESULTS

We evaluated the Guide-LLM by testing its ability to navigate through a simulated indoor environment based on user queries. The simulations are divided into scenarios designed to highlight distinct aspects of the Guide-LLM’s utility in navigation, decision-making, error handling, and hazard detection and personalization.

A. Experiment Setup

We used iGibson [52] simulator to validate Guide-LLM. The simulator was deployed using AWS G4dn.2xlarge instance and LLM models were deployed in the cloud using API.

TABLE III

LOCALIZATION ERROR DETECTION AND RECOVERY

Error Scenario	Success Rate (%)
Localization Error Detection	90%
Recovery	72%

TABLE IV

HAZARD DETECTION RESULTS

Scenario	# Trials	TP	FP	FN	TN
Potential Hazard	30	25		5	
Non-Potential Hazard	30		10		20

B. Model Selection and Topological Navigation

To assess different LLM’s ability to reach specified destinations and select optimal LLM for Guide-LLM, we performed evaluations in two indoor environments within the iGibson simulator [49]. These environments included a three-room house and a larger office-like setting with multiple hallways and offices, as shown in Fig. 3. In both environments, the agent needed to reach a specified destination using four key components: (1) a user query, (2) a text-based topological map, (3) a vector database, and (4) its own reasoning ability. To examine the impact of model scale on navigation performance, we selected five vision-enabled LLMs: three models with over 90 billion parameters and two models with fewer than 11 billion parameters. We repeated each trial 50 times per environment to ensure statistical robustness. If the agent successfully arrived at the destination, we continued issuing new navigation queries until it eventually failed to reach a subsequent destination. If the agent failed to reach its current destination, we reset the robot’s position and counted that attempt as a failure. Table II shows that GPT-4o[48] achieved the highest overall success rate. Most navigation failures resulted from the LLM “forgetting” which command it should generate or confused with turning directions, suggesting that limited real-world grounding can hinder consistent decision-making. Furthermore, our experiments reveal that model size plays a crucial role in navigation success rates: GPT-4o’s performance was approximately three times higher than that of Llama 3.2 Vision 11B, suggesting a positive correlation between parameter count and navigation capability.

C. Localization Error Detection and Recovery

To evaluate Guide-LLM’s resilience against localization failures, we introduced the ‘kidnapped robot’ scenario in the office environment. During navigation, the robot was randomly teleported to a different node, forcing the model to detect error and recover from it. We performed this evaluation 50 times in total, ignoring naturally failed scenarios where navigation failed due to LLM’s navigation failure. If the agent successfully recognized the mismatch between its expected location and current observation, we considered the localization error detection successful. If the agent re-localize and reached the goal by updating its plan based on the new location, the trial was marked as a successful recovery. Table III shows the localization error detection and recovery

rates. In addition, we observed that most failures occurred in feature-sparse corridors or visually similar places. We also observe that recovery failures happen when the Guide-LLM recognized the localization error but carries on with the previous path.

D. Commonsense Reasoning for Hazard Detection

The objective of this experiment was to evaluate the agent’s ability to identify potential hazards in the environment, communicate these risks to the user, and adapt its actions based on user decisions. To test this capability, we placed obstacles and potential hazards (warning signs, physical barriers) and non-potential hazards along the navigation path (chair, pot). As shown in Table IV, we conducted 60 trials: 30 with objects and hazards with the potential to impede navigation, while the other 30 with non-hazardous objects that did not interfere with safe travel. In the hazard scenario, we recorded 25 true positives (TP), where the agent correctly identified a hazard and warned the user, and 5 false negatives (FN), where it failed to recognize a hazard and proceeded without warning. These false negatives typically resulted from limited height awareness of both the Guide-LLM and the user. In the non-hazard scenario, we observed 10 false positives (FP), where the agent mistakenly flagged hazard, and 20 true negatives (TN), where it ignored a harmless item. Figure 3 illustrates a typical hazard scenario in which the agent demonstrates commonsense reasoning beyond navigation. Upon detecting a hazard, it alerts the user and proposes an alternate route, underscoring the system’s capacity to dynamically adapt when faced with unexpected obstacles.

E. Personalization Potential

Prior research on navigational aids indicates that personalization is vital for enhancing user experience [53]. As shown in Table I, current navigation aids are limited in their ability to incorporate personalized preferences. In contrast, Guide-LLM supports diverse user needs through natural language interaction, enabling users to easily communicate specific requirements. To illustrate how user preferences can be addressed, we renamed certain spaces in the text-based topological map (e.g., “concert hall,” “food court,” “noisy area,” “quiet area”). We then conducted 50 repeated trials on the same route, applying various user preferences and phrasings (e.g., avoiding stairs, noisy areas, or avoiding places likely to be crowded). In each trial, Guide-LLM adapted its path planning accordingly, even selecting routes up to three times longer if it aligned with the user’s stated requirements. This is possible by the integration of the text-based map with the path planning module, allowing the Guide-LLM to evaluate alternative routes and plan ahead. Some failures were seen when the number of possible routes was excessively large, potentially exceeding the maximum context window and leading to suboptimal decisions. This test resulted 92% success rate. Additionally, we evaluated Guide-LLM’s ability to adjust speed by providing multiple instructional cues (e.g., asking it to speed up or slow down) to

see whether the system could infer user intent. We observed that Guide-LLM consistently increased or decreased walking speed without any failures. For characteristic changes such as providing detailed, step-by-step navigation or shifting conversational tone (e.g., formal, casual, or informal), the system met user expectations in 43 out of 50 trials.

F. LLM Guided Navigation Ablation Study

We conducted an ablation study to examine how each core component of the Guide-LLM framework contributes to navigation performance. As shown in Table II, the full system achieved an 86% success rate when navigating through an office environment. We then systematically removed individual components and measured the resulting performance drops. Removing the system prompt reduced the success rate to 0%, indicating that prompt-driven context is essential for the agent’s high-level understanding and instruction formulation. Next, we removed the path planning module and directly supplied the text-based map to the agent. This change lowered the success rate to 40%. The primary reason is that adding a large text map to the system prompt can exceed the context window, which degrades the agent’s reasoning capability due to processing bottlenecks associated with large prompts [54], [55], [56]. Finally, removing the image retrieval system forced users to specify the starting location, thereby altering the experimental conditions and increasing user burden. Since it no longer tested the agent’s ability to visually infer starting points, we excluded this configuration from our comparative analysis to maintain fairness.

V. CONCLUSION AND FUTURE WORK

We introduce Guide-LLM, an embodied large language model framework that can provide efficient, adaptive, and personalized navigation assistance for people with visual impairments. By combining a text-based topological map with common sense reasoning, our system reduces the need for explicit user instructions, enabling personalized navigation, offering a more intuitive navigational experience. Simulation results demonstrate that Guide-LLM excels in tasks such as localization error detection recovery, hazard detection, and personalization. As this research focused on developing and validating the technical feasibility of LLM based navigation aid, we plan to focus on the following directions in the future: Optimizing smaller language models to match or approximate the performance of larger ones, allowing on-device processing without reliance on cloud-based services. Expanding the text-based map creation process to autonomously generate and update environmental representations in large, complex buildings such as multi-story shopping center. Attaching depth camera to provide depth information to improve on hazard detection. Conducting real-world trials with visually impaired participants to evaluate usability, comfort, and overall effectiveness in navigation tasks. By addressing these challenges, we aim to refine Guide-LLM into a comprehensive assistive solution that empowers visually impaired individuals with more flexible, accessible, and safe navigation.

REFERENCES

- [1] D. L. Rudman and M. Durdle, "Living with fear: The lived experience of community mobility among older adults with low vision," *Journal of aging and physical activity*, vol. 17, no. 1, pp. 106–122, 2008.
- [2] N. A. Giudice and G. E. Legge, "Blind navigation and the role of technology," *The engineering handbook of smart technology for aging, disability, and independence*, pp. 479–500, 2008.
- [3] M. Y. Wang, J. Rousseau, H. Boisjoly, H. Schmaltz, M.-J. Kergoat, S. Moghadaszadeh, F. Djafari, and E. E. Freeman, "Activity limitation due to a fear of falling in older adults with eye disease," *Investigative ophthalmology & visual science*, vol. 53, no. 13, pp. 7967–7972, 2012.
- [4] W. Jeamwathanachai, M. Wald, and G. Wills, "Indoor navigation by blind people: Behaviors and challenges in unfamiliar spaces and buildings," *British Journal of Visual Impairment*, vol. 37, no. 2, pp. 140–153, 2019.
- [5] Z. Başıoğlu, J. Gualtieri, M. T. Sachs, and E. A. Cooper, "Navigational aid use by individuals with visual impairments," in *Journal on technology and persons with disabilities... Annual International Technology and Persons with Disabilities Conference*, vol. 8. NIH Public Access, 2020, p. 22.
- [6] N. A. Giudice, B. A. Guenther, T. M. Kaplan, S. M. Anderson, R. J. Knuesel, and J. F. Cioffi, "Use of an indoor navigation system by sighted and blind travelers: Performance similarities across visual status and age," *ACM Transactions on Accessible Computing (TACCESS)*, vol. 13, no. 3, pp. 1–27, 2020.
- [7] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat *et al.*, "Gpt-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.
- [8] A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Yang, A. Fan *et al.*, "The llama 3 herd of models," *arXiv preprint arXiv:2407.21783*, 2024.
- [9] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," *Advances in neural information processing systems*, vol. 36, 2024.
- [10] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PMLR, 2021, pp. 8748–8763.
- [11] J. Li, D. Li, S. Savarese, and S. Hoi, "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," in *International conference on machine learning*. PMLR, 2023, pp. 19730–19742.
- [12] T. Williams, C. Matuszek, R. Mead, and N. Depalma, "Scarecrows in Oz: The Use of Large Language Models in HRI," *ACM Transactions on Human-Robot Interaction*, vol. 13, no. 1, pp. 1–11, 1 2024.
- [13] K. Rana, J. Haviland, S. Garg, J. Abou-Chakra, I. Reid, and N. Suennderhauf, "Sayplan: Grounding large language models using 3d scene graphs for scalable robot task planning," in *7th Annual Conference on Robot Learning*, 2023.
- [14] X. Jiang, Y. Dong, L. Wang, F. Zheng, Q. Shang, G. Li, Z. Jin, and W. Jiao, "Self-planning Code Generation with Large Language Models," *ACM Transactions on Software Engineering and Methodology*, 6 2024.
- [15] D. Shah, B. Osiński, S. Levine *et al.*, "Lm-nav: Robotic navigation with large pre-trained models of language, vision, and action," in *Conference on robot learning*. PMLR, 2023, pp. 492–504.
- [16] Y. Cui, S. Huang, J. Zhong, Z. Liu, Y. Wang, C. Sun, B. Li, X. Wang, and A. Khajepour, "Drivellm: Charting the Path Toward Full Autonomous Driving With Large Language Models," *IEEE Transactions on Intelligent Vehicles*, vol. 9, no. 1, pp. 1450–1464, 1 2024.
- [17] S. Liu, A. Hasan, K. Hong, R. Wang, P. Chang, Z. Mizrachi, J. Lin, D. L. McPherson, W. A. Rogers, and K. Driggs-Campbell, "Dragon: A dialogue-based robot for assistive navigation with visual language grounding," *IEEE Robotics and Automation Letters*, 2024.
- [18] C. Huang, O. Mees, A. Zeng, and W. Burgard, "Visual language maps for robot navigation," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2023, pp. 10608–10615.
- [19] S. Chen, X. Chen, C. Zhang, M. Li, G. Yu, H. Fei, H. Zhu, J. Fan, and T. Chen, "LI3da: Visual interactive instruction tuning for omni-3d understanding reasoning and planning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 26428–26438.
- [20] S. Yang, J. Liu, R. Zhang, M. Pan, Z. Guo, X. Li, Z. Chen, P. Gao, Y. Guo, and S. Zhang, "Lidar-llm: Exploring the potential of large language models for 3d lidar understanding," *arXiv preprint arXiv:2312.14074*, 2023.
- [21] Y. Hong, H. Zhen, P. Chen, S. Zheng, Y. Du, Z. Chen, and C. Gan, "3d-llm: Injecting the 3d world into large language models," *Advances in Neural Information Processing Systems*, vol. 36, pp. 20482–20494, 2023.
- [22] R. Xu, X. Wang, T. Wang, Y. Chen, J. Pang, and D. Lin, "Pointllm: Empowering large language models to understand point clouds," *arXiv preprint arXiv:2308.16911*, 2023.
- [23] M. Swobodzinski and M. Raubal, "An indoor routing algorithm for the blind: development and comparison to a routing algorithm for the sighted," *International Journal of Geographical Information Science*, vol. 23, no. 10, pp. 1315–1343, 2009.
- [24] S. Shafiqe, W. Setti, C. Campus, S. Zanchi, A. Del Bue, and M. Gori, "How path integration abilities of blind people change in different exploration conditions," *Frontiers in Neuroscience*, vol. 18, p. 1375225, 2024.
- [25] G. Zhou, Y. Hong, and Q. Wu, "Navgpt: Explicit reasoning in vision-and-language navigation with large language models," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 7, 2024, pp. 7641–7649.
- [26] P. Slade, A. Tambe, and M. J. Kochenderfer, "Multimodal sensing and intuitive steering assistance improve navigation and mobility for people with impaired vision," *Science Robotics*, vol. 6, no. 59, 10 2021.
- [27] M. M. Islam, M. S. Sadi, and T. Bräunl, "Automated walking guide to enhance the mobility of visually impaired people," *IEEE Transactions on Medical Robotics and Bionics*, vol. 2, no. 3, pp. 485–496, 2020.
- [28] L. Jin, H. Zhang, and C. Ye, "A wearable robotic device for assistive navigation and object manipulation," in *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2021, pp. 765–770.
- [29] Y. Bouteraa, "Design and development of a wearable assistive device integrating a fuzzy decision support system for blind and visually impaired people," *Micromachines*, vol. 12, no. 9, p. 1082, 2021.
- [30] G. Li, J. Xu, Z. Li, C. Chen, and Z. Kan, "Sensing and navigation of wearable assistance cognitive systems for the visually impaired," *IEEE Transactions on Cognitive and Developmental Systems*, vol. 15, no. 1, pp. 122–133, 2022.
- [31] X. Liu, B. Wang, and Z. Li, "Vision-based wearable steering assistance for people with impaired vision in jogging," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2024, pp. 15270–15275.
- [32] S. Yao, J. Zhao, D. Yu, N. Du, I. Shafran, K. Narasimhan, and Y. Cao, "React: Synergizing reasoning and acting in language models," *arXiv preprint arXiv:2210.03629*, 2022.
- [33] P. Sermanet, T. Ding, J. Zhao, F. Xia, D. Dwibedi, K. Gopalakrishnan, C. Chan, G. Dulac-Arnold, S. Maddineni, N. J. Joshi *et al.*, "Robovqa: Multimodal long-horizon reasoning for robotics," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2024, pp. 645–652.

- [34] D. Shah, M. R. Equi, B. Osiński, F. Xia, B. Ichter, and S. Levine, "Navigation with large language models: Semantic guesswork as a heuristic for planning," in *Conference on Robot Learning*. PMLR, 2023, pp. 2683–2699.
- [35] A. Rajvanshi, K. Sikka, X. Lin, B. Lee, H.-P. Chiu, and A. Velasquez, "Saynav: Grounding large language models for dynamic planning to navigation in new environments," in *Proceedings of the International Conference on Automated Planning and Scheduling*, vol. 34, 2024, pp. 464–474.
- [36] W. Huang, P. Abbeel, D. Pathak, and I. Mordatch, "Language models as zero-shot planners: Extracting actionable knowledge for embodied agents," in *International conference on machine learning*. PMLR, 2022, pp. 9118–9147.
- [37] J. Chen, B. Lin, R. Xu, Z. Chai, X. Liang, and K.-Y. Wong, "Mapgpt: Map-guided prompting with adaptive path planning for vision-and-language navigation," in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2024, pp. 9796–9810.
- [38] V. S. Dorbala, J. F. Mullen, and D. Manocha, "Can an Embodied Agent Find Your "Cat-shaped Mug"? LLM-Based Zero-Shot Object Navigation," *IEEE Robotics and Automation Letters*, vol. 9, no. 5, pp. 4083–4090, 5 2024.
- [39] D. Fried, R. Hu, V. Cirik, A. Rohrbach, J. Andreas, L.-P. Morency, T. Berg-Kirkpatrick, K. Saenko, D. Klein, and T. Darrell, "Speaker-follower models for vision-and-language navigation," *Advances in neural information processing systems*, vol. 31, 2018.
- [40] C.-Y. Ma, J. Lu, Z. Wu, G. AlRegib, Z. Kira, R. Socher, and C. Xiong, "Self-monitoring navigation agent via auxiliary progress estimation," *arXiv preprint arXiv:1901.03035*, 2019.
- [41] X. Li, A. Yuan, and X. Lu, "Vision-to-language tasks based on attributes and attention mechanism," *IEEE transactions on cybernetics*, vol. 51, no. 2, pp. 913–926, 2019.
- [42] A. B. Vasudevan, D. Dai, and L. Van Gool, "Talk2nav: Long-range vision-and-language navigation with dual attention and spatial memory," *International Journal of Computer Vision*, vol. 129, pp. 246–266, 2021.
- [43] X. Wang, W. Xiong, H. Wang, and W. Y. Wang, "Look before you leap: Bridging model-free and model-based reinforcement learning for planned-ahead vision-and-language navigation," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 37–53.
- [44] H. Wang, Q. Wu, and C. Shen, "Soft expert reward learning for vision-and-language navigation," in *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16*. Springer, 2020, pp. 126–141.
- [45] S. Garg, K. Rana, M. Hosseinzadeh, L. Mares, N. Sünderhauf, F. Dayoub, and I. Reid, "Robohop: Segment-based topological map representation for open-world visual navigation," *arXiv preprint arXiv:2405.05792*, 2024.
- [46] R. Schumann, W. Zhu, W. Feng, T.-J. Fu, S. Riezler, and W. Y. Wang, "Velma: Verbalization embodiment of llm agents for vision and language navigation in street view," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 17, 2024, pp. 18924–18933.
- [47] R. V. Jawale, M. V. Kadam, R. S. Gaikawad, and L. S. Kondaka, "Ultrasonic navigation based blind aid for the visually impaired," in *2017 IEEE International Conference on Power, Control, Signals and Instrumentation Engineering (ICPCSI)*. IEEE, 2017, pp. 923–928.
- [48] OpenAI, "Hello gpt-4," <https://openai.com/index/hello-gpt-4/>, accessed: Sept. 14, 2024.
- [49] Anthropic, "Introducing claude 3.5 sonnet," <https://www.anthropic.com/news/claude-3-5-sonnet>, June 2024, accessed: February 26, 2025.
- [50] M. AI, "Llama 3.2: Revolutionizing edge ai and vision with open, customizable models," <https://ai.meta.com/blog/llama-3-2-connect-2024-vision-edge-mobile-devices/>, September 2024.
- [51] S. Bai, K. Chen, X. Liu, J. Wang, W. Ge, S. Song, K. Dang, P. Wang, S. Wang, J. Tang, H. Zhong, Y. Zhu, M. Yang, Z. Li, J. Wan, P. Wang, W. Ding, Z. Fu, Y. Xu, J. Ye, X. Zhang, T. Xie, Z. Cheng, H. Zhong, Z. Yang, H. Xu, J. Lin *et al.*, "Qwen2.5-vl technical report," *arXiv preprint arXiv:2502.13923*, February 2025. [Online]. Available: <https://arxiv.org/abs/2502.13923>
- [52] C. Li, F. Xia, R. Martín-Martín, M. Lingelbach, S. Srivastava, B. Shen, K. Vainio, C. Gokmen, G. Dharan, T. Jain *et al.*, "igibson 2.0: Object-centric simulation for robot learning of everyday household tasks," *arXiv preprint arXiv:2108.03272*, 2021.
- [53] D. Ahmetovic, J. Guerreiro, E. Ohn-Bar, K. M. Kitani, and C. Asakawa, "Impact of expertise on interaction preferences for navigation assistance of visually impaired individuals," in *Proceedings of the 16th International Web for All Conference*, 2019, pp. 1–9.
- [54] C. An, S. Gong, M. Zhong, X. Zhao, M. Li, J. Zhang, L. Kong, and X. Qiu, "L-eval: Instituting standardized evaluation for long context language models," *arXiv preprint arXiv:2307.11088*, 2023.
- [55] Y. Bai, X. Lv, J. Zhang, H. Lyu, J. Tang, Z. Huang, Z. Du, X. Liu, A. Zeng, L. Hou *et al.*, "Longbench: A bilingual, multitask benchmark for long context understanding," *arXiv preprint arXiv:2308.14508*, 2023.
- [56] M. Levy, A. Jacoby, and Y. Goldberg, "Same task, more tokens: the impact of input length on the reasoning performance of large language models," *arXiv preprint arXiv:2402.14848*, 2024.