Generating Contextually-Relevant Navigation Instructions for Blind and Low Vision People

Zain Merchant[1] Abrar Anwar[1] Emily Wang[1] Souti Chattopadhyay[1] Jesse Thomason[1]

Abstract— **Navigating unfamiliar environments presents significant challenges for blind and low-vision (BLV) individuals. In this work, we construct a dataset of images and goals across different scenarios such as kitchens or outdoor navigation. We then investigate how grounded instruction generation methods can provide contextually-relevant navigational guidance to users in these instances. Through a study involving sighted users, we demonstrate that large pretrained language models can produce correct and useful instructions perceived as beneficial for BLV users. We also conduct a survey and interview with 4 BLV users and observe useful insights on preferences for different instructions based on the scenario.**

I. INTRODUCTION AND BACKGROUND

Nearly 253 million people struggle with visual impairment worldwide, where 36 million of these individuals are blind [1]. Dealing with the complexities of daily life poses significant challenges for these individuals, particularly when exploring unfamiliar environments. Traditional aids such as canes and guide dogs are vital in facilitating mobility and independence. However, these tools have limitations in conveying the rich visual information that sighted individuals rely on for navigation and object recognition.

There has been recent growth on using vision-and-language models as visual assistants that can interactively communicate with a user to provide feedback [2]. Additionally, after interviews with blind and low vision individuals, prior work has noted a critical issue with the use of guide dogs: the communication from the user to the dog is unimodal. However, there may be questions users want to ask, such as “Is it safe to cross the street?” [3]. They posit that any robotic guide dog should handle complex interaction with a user to answer these types of questions. As the companies and the research community begin to integrate large language models into these robot systems, it is important to understand the role of a language model’s contextual understanding capabilities in providing personalized, informative feedback tailored to a user’s specific goals and surroundings. In this work, we focus on navigation assistance and investigate the usefulness and contextual relevance of generated instructions for navigational assistance using large language models (LLMs) and vision-and-language models (VLMs).

Existing work in the field of blind and low vision (BLV) navigation assistance has primarily focused on lower-level navigation tasks such as obstacle avoidance [4]–[8]. Many existing systems gather different kinds of information from the environment, using basic object detection [9] often combined with auditory output [10] based on templates [11] to generate descriptions or simply list objects [12]. While these approaches offer valuable assistance, they often overlook the importance of context and relevance in delivering instructions to BLV users [10], [13].

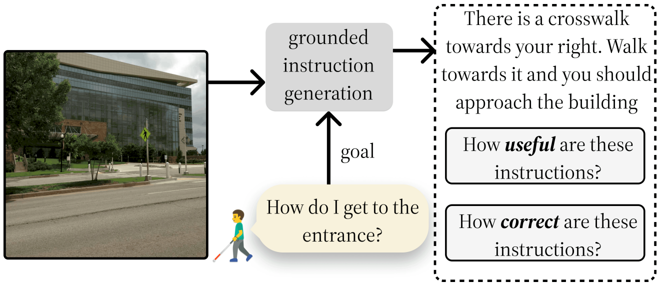
grounded instruction generation
goal
How do I get to the entrance?
There is a crosswalk towards your right. Walk towards it and you should approach the building
How ***useful*** are these instructions?
How ***correct*** are these instructions?

Fig. 1: We formulate the problem of providing contextually-relevant navigational instructions to blind and low vision (BLV) people as a grounded instruction generation task, which we then evaluate with sighted and BLV participants in a user study.

Descriptions are subjective and depend on the user’s context [10], [15], and overly generic or unnecessary information is often not preferred by users [13], [16], [17]. Noting these flaws with prior work and considering recent work on the importance of context for BLV participants [10], [13], we propose to augment such systems with LLMs and VLMs and understand whether these methods are able to generate contextually-relevant instructions.

II. PROBLEM SETTING

We want to generate instructions that are *useful* and goal-aware for BLV users, as shown in Figure 1. We frame this problem as a grounded instruction generation task, where a model $S(w_0,...,w_N|g,o)$ uses an egocentric image to generate an instruction $w=[w_0,...,w_N]$ that describes a route for an BLV user to reach a goal g given an image observation o. The goal g is semantic task context from the user like “How do I get to the building’s entrance?” The amount of objects and semantic information that can be conveyed to a user based on an image is often innumerable; however, only a subset of this information is useful to the user to accomplish their goal. In this work, we are interested in *how the environment and goal impact the usefulness of the generated instruction.*

VizWiz [14] is a collection of photographs captured by BLV participants and serves as a basis for formulating tasks related to object detection and visual question-answering.

[1] Zain Merchant, Abrar Anwar, Emily Wang, Souti Chattopadhyay, and Jesse Thomason are with the Thomas Lord Department of Computer Science, University of Southern California, Los Angeles, CA, USA
Contact: zsmercha@usc.edu

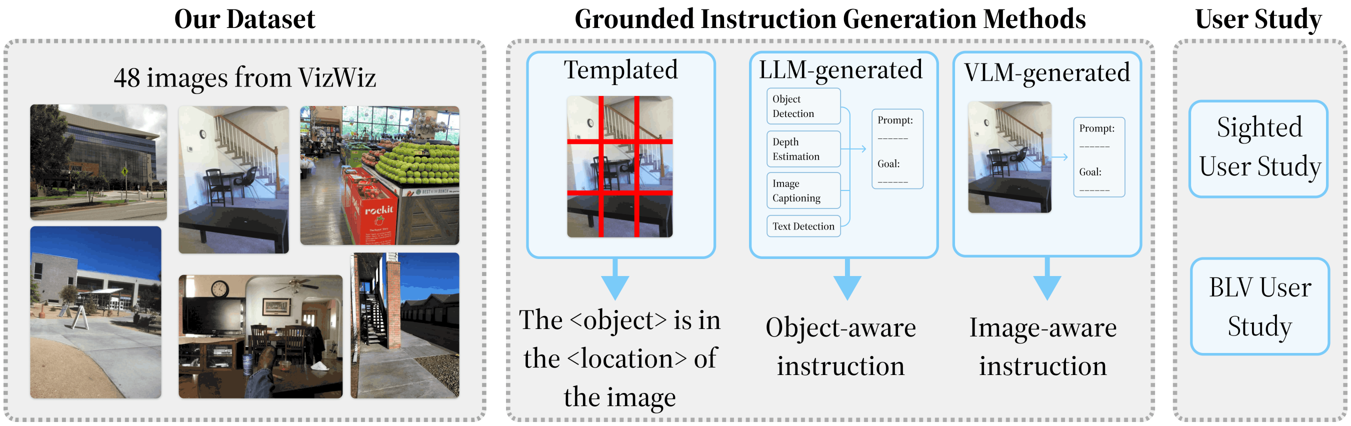

Fig. 2: *Left:* We select 48 images from indoor and outdoor environments in VizWiz [14] and annotate them with navigation goals. *Middle:* We design three instruction generation methods, described further in Section III. *Right:* These generated instructions are then evaluated in a user study with sighted and BLV participants.

Since VizWiz does not consider navigation instruction generation, we created our own dataset by selecting 48 images from VizWiz that were relevant to navigation from four different environments: offices, kitchens, general indoor, and general outdoor settings. We then assigned each image a goal, such as “Where is the TV remote”, which in navigation tasks, may not be visible in a single frame due to the object’s size or placement.

In our setting, we are considering only a single image, but in practical applications, the image alone may not contain the answer to the goal. For example, in the case of ”Where is the TV remote”, the remote may not be visible and may require providing potential hints. To capture these kinds of problems, we construct a split such that 20% of our dataset are **Hard** examples which require a model to reason more extensively on how to provide instructions for the goal. The complementary split is referred to as the **Easy** split. Each of the images was annotated by sighted volunteers with instructions that would solve the goal.

III. Grounded Instruction Generation

We compare methods for generating navigation instructions from single images paired with semantic task context.

Human instructions. As a point of reference for machine-generated instructions, we sent selected image-goal pairs to four human annotators tasked with generating instructions for BLV users.

Template Instructions. Past work [11] extracted the object of interest from the user’s goal query. For example, for the sentence “where is the textbook”, this baseline extracts the object “textbook”, then detects where the object is using the OWL-ViT open-vocabulary object detector [18]. This method then localizes the object into nine predefined areas (top left, center, bottom right, etc.). If a user asks about the location of a microwave, the system can concisely respond: “The microwave is at the top left.”

LLM-generated Instructions. Text-only LLMs have good commonsense reasoning abilities given text input; however, unlike VLMs, they cannot take images as input. In order to generate grounded navigation instructions, we take a Socratic model [19] approach, where we provide the outputs of various off-the-shelf models as additional information for the LLM. We use off-the-shelf object detection [18], depth estimation [20], image captioning, and optical character recognition. These inputs are formatted into a prompt with the goal and a few in-context examples. We use GPT4 [21] as the LLM and get contextually relevant, grounded instructions as the output.

VLM-generated Instructions. VLMs are able to take images as input, so rather than a Socratic models approach, we use GPT-4 Vision [21] with the image as an input and prompt the model to generate an instruction.

IV. Study Design

We conducted two IRB-approved human subject studies with sighted and BLV participants.

A. Sighted User Study

Sighted participants were asked to take a survey to rate instructions given an image and a goal. They rated the instructions in terms of *Correctness* and *Usefulness* on a 1 to 7 Likert scale based on the following definition. We define *Correctness* as how accurate the instruction is with respect to directions and objects in the image. This metric also verifies the instruction generation methods generate accurate instructions. We define *Usefulness* as how useful the instruction would be to a BLV user to help achieve their goal, considering its relevance and safety. We recruited eight sighted participants, who each viewed six images across four methods.

B. BLV User Study

Due to the low-incidence of the BLV population, we recruited three blind and one low-vision participant. Similar to the sighted user study, each image is rated by a BLV participant for *Usefulness*. We do not collect *Correctness* ratings as it depends on the level of visual impairment and how they perceive the scene. After the survey, we conducted a semi-structured interview. Survey questions were aimed to elicit their thoughts on navigating different environments, including social spaces, and their thoughts about the kinds of methods they experienced.

V. Results

We present our sighted and BLV user study results.

A. Sighted Survey Results

With our sighted user survey, we find that users find the generated instructions correct and useful, which shows promise for these methods to be tested with BLV users.

Generated instructions are similarly useful to human annotated instructions. Table I shows the aggregated correctness and usefulness scores given by sighted users across methods. Users found human-generated instructions more accurate than LLM- and VLM-generated instructions. In contrast, the difference between usefulness ratings between human, LLM-, and VLM-generated instructions was much smaller. The difference in usefulness ratings between the VLM and LLM could be explained by harder-to-answer instances benefiting from the input of the entire image, while easy-to-answer instances can be solved more directly with object detectors, as supported by Table II.

Users find different amounts of usefulness of instructions depending on the environment. Figure 3 shows box plots of the usefulness scores of sighted users across different environments. We find that the VLM and human-generated instructions have similar usefulness score distributions compared to the LLM-generated and templated instructions. We also observe that instructions generated for office and general indoor environments are rated more useful than kitchen and outdoor instructions. This trend could be because the instruction generation methods have to reason about more complex scenes, or users having different expectations in these scenes.

	Correctness	**Usefulness**
Templated	3.85 ± 1.95	1.96 ± 1.05
LLM-based	4.73 ± 1.61	4.24 ± 1.38
VLM-based	4.75 ± 1.78	4.52 ± 1.70
Human	5.46 ± 1.46	4.73 ± 1.83

TABLE I: Sighted user ratings for **Correctness** and **Usefulness**. Human-annotated instructions are more accurate compared to the other methods, but the LLM- and VLM-generated instructions were rated similarly useful.

	Easy Split	**Hard** Split
Templated	2.11 ± 1.09	1.40 ± 0.70
LLM-based	4.39 ± 1.75	5.00 ± 1.49
VLM-based	4.20 ± 1.30	4.40 ± 1.71
Human	4.58 ± 1.80	5.30 ± 1.95

TABLE II: **Usefulness** scores for the **Easy** and **Hard** splits from the 48 image-goal pairs. Interestingly, the **Easy** split was rated lower than the **Hard** split.

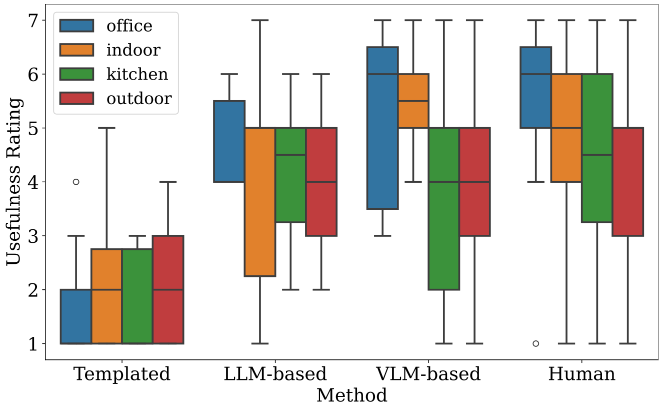

Fig. 3: Sighted participant **Usefulness** ratings over the generated instructions for 48 image-goal pairs across four methods separated by environment. VLM-based instructions had similar ratings across environments to humans. The LLM-based model was rated slightly less useful.

B. BLV Survey Results

The sighted user results indicate that the generated instructions are correct and show trends about the role of context. Though these results provide insights about the generated instructions, we focus on the quantitative and qualitative results from our BLV user study.

BLV participants rate methods as less useful compared to sighted participants. As shown in Table III, we find the LLM- and VLM-generated instructions were rated slightly lower to the human. Unsurprisingly, the template instructions were consistently not useful. Due to our small size of our BLV user study, we will focus primarily on the qualitative semi-structured interview in the next section.

C. BLV Qualitative Interview

Different environments change preferences on the kinds of instructions. Participants indicated that in confined spaces such as a kitchen, they have a preference for less broad instructions, whereas outdoor, open spaces can be more broad (e.g. "walk forward until you reach the corner"). The low-vision participant noted that lighting and audio cues could provide a means for useful guidance.

Generative methods rely on visual cues. Participants noted that some responses relied on visual cues such as "it's near the big sign" which is not useful. Their suggestion was to make the system more specific and to focus on integrating more spatial awareness as those instructions are most useful.

Participants preferred specific directions (e.g. "take a few steps to your right", "10 degrees to the right") over vague ones (e.g. "the table is in the center"). However, they noted that specificity is not always useful. For example, knowing how many objects are on a table might be too much information to be given at once and could be asked as a separate question.

What makes a useful navigation assistant? Several participants indicated that an ideal system would tell them where things are laid out in relation to other things that they

Method	**Usefulness**
Templated	2.00 ± 1.29
LLM-based	3.97 ± 1.78
VLM-based	4.45 ± 1.73
Human	4.03 ± 1.80

TABLE III: **Usefulness** ratings from our BLV participants. The VLM-based instructions were rated as more useful than all of instructions.

can reason about. For example, “the sponge is in the bottom right of the basin” is helpful, but “the bench is to the right of the sign” is not. In contrast to prior work in this space and systems like Google Maps, participants noted frustration with instructions that stated a precise number of feet to walk, especially since these systems cannot tell the user when they have reached that distance. Instructions like “walk until you reach the corner” would resolve this issue. Thus, leveraging the relationship between the goals with one’s surroundings can be helpful. One participant found the LM-generated instructions to be wordy or condescending, motivating investigations into preferences in *how* these models communicate information.

VI. Conclusion, Ethics, and Limitations

LLM and VLM-based methods for grounded instruction generation show great promise in integrating with assistive technologies. However, a significant challenge associated with using these models is their tendency to produce hallucinations or inaccurate generations. Poorly generated instructions can lead to confusion and put users in potentially hazardous situations.

We also recognize the reference instructions written by sighted annotators may not be tailored to how a BLV user may want to be given instructions, as the annotators were not expertly trained to communicate with BLV users.

LLMs and VLMs are also susceptible to biases present in their training data [22]. It’s important to ensure these technologies are trained on diverse data sets that accurately represent the variety of cultures and environments that may be encountered so that these assistive technologies can serve users in an equitable manner. By emphasizing the role of context in the generation of instructions for BLV users, we hope our work can initiate a community discussion on how to handle the many possible scenarios a user could experience.

References

[1] P. Ackland, S. Resnikoff, and R. Bourne, “World blindness and visual impairment: despite many successes, the problem is growing,” *Community eye health*, 2017.

[2] Be My Eyes, “Be my eyes,” https://www.bemyeyes.com/.

[3] H. Hwang, H.-T. Jung, N. A. Giudice, J. Biswas, S. I. Lee, and D. Kim, “Towards robotic companions: Understanding handler-guide dog interactions for informed guide dog robot design,” *Conference on Human Factors in Computing Systems (CHI)*, 2024.

[4] S. Real and A. Araujo, “Navigation systems for the blind and visually impaired: Past work, challenges, and open problems,” *Sensors*, 2019. [Online]. Available: https://www.mdpi.com/1424-8220/19/15/3404

[5] A. Cassinelli, C. Reynolds, and M. Ishikawa, “Augmenting spatial awareness with haptic radar,” in *IEEE International Symposium on Wearable Computers*, 2006.

[6] R. N. Kandalan and K. Namuduri, “Techniques for constructing indoor navigation systems for the visually impaired: A review,” *IEEE Transactions on Human-Machine Systems*, 2020.

[7] Y. Lin, K. Wang, W. Yi, and S. Lian, “Deep learning based wearable assistive system for visually impaired people,” in *ICCV Workshops*, Oct 2019.

[8] Z. Bauer, A. Dominguez, E. Cruz, F. Gomez-Donoso, S. Orts-Escolano, and M. Cazorla, “Enhancing perception for the visually impaired with deep learning techniques and low-cost wearable sensors,” *Pattern Recognition Letters*, 2020.

[9] M. Leo, A. Furnari, G. G. Medioni, M. Trivedi, and G. M. Farinella, “Deep learning for assistive computer vision,” in *ECCV Workshops*, September 2018.

[10] H. Walle, C. De Runz, B. Serres, and G. Venturini, “A survey on recent advances in ai and vision-based methods for helping and guiding visually impaired people,” *Applied Sciences*, 2022. [Online]. Available: https://www.mdpi.com/2076-3417/12/5/2308

[11] K. Thakoor, N. Mante, C. Zhang, C. Siagian, J. Weiland, L. Itti, and G. Medioni, “A system for assisting the visually impaired in localization and grasp of desired objects,” in *ECCV Workshops*, L. Agapito, M. M. Bronstein, and C. Rother, Eds., 2014.

[12] S. Malek, F. Melgani, M. L. Mekhalfi, and Y. Bazi, “Real-time indoor scene description for the visually impaired using autoencoder fusion strategies with visible cameras,” *Sensors*, 2017. [Online]. Available: https://www.mdpi.com/1424-8220/17/11/2641

[13] K. M. P. Hoogsteen, S. Szpiro, G. Kreiman, and E. Peli, “Beyond the cane: Describing urban scenes to blind people for mobility tasks,” *ACM Trans. Access. Comput.*, 2022. [Online]. Available: https://doi.org/10.1145/3522757

[14] D. Gurari, Q. Li, A. J. Stangl, A. Guo, C. Lin, K. Grauman, J. Luo, and J. P. Bigham, “Vizwiz grand challenge: Answering visual questions from blind people,” in *CVPR*, 2018.

[15] E. Kreiss, C. Bennett, S. Hooshmand, E. Zelikman, M. R. Morris, and C. Potts, “Context matters for image descriptions for accessibility: Challenges for referenceless evaluation metrics,” *EMNLP*, 2022.

[16] M. K. Scheuerman, W. Easley, A. Abdolrahmani, A. Hurst, and S. Branham, “Learning the language: The importance of studying written directions in designing navigational technologies for the blind,” in *CHI Extended Abstracts on Human Factors in Computing Systems*, 2017. [Online]. Available: https://doi.org/10.1145/3027063.3053260

[17] M. A. Williams, C. Galbraith, S. K. Kane, and A. Hurst, “”just let the cane hit it”: how the blind and sighted see navigation differently,” in *ACM SIGACCESS Conference on Computers & Accessibility*, ser. ASSETS, 2014. [Online]. Available: https://doi.org/10.1145/2661334.2661380

[18] M. Minderer, A. Gritsenko, A. Stone, M. Neumann, D. Weissenborn, A. Dosovitskiy, A. Mahendran, A. Arnab, M. Dehghani, Z. Shen *et al.*, “Simple open-vocabulary object detection,” in *ECCV*, 2022.

[19] A. Zeng, M. Attarian, B. Ichter, K. Choromanski, A. Wong, S. Welker, F. Tombari, A. Purohit, M. Ryoo, V. Sindhwani, J. Lee, V. Vanhoucke, and P. Florence, “Socratic models: Composing zero-shot multimodal reasoning with language,” 2022.

[20] D. Kim, W. Ga, P. Ahn, D. Joo, S. Chun, and J. Kim, “Global-local path networks for monocular depth estimation with vertical cutdepth,” *CoRR*, vol. abs/2201.07436, 2022. [Online]. Available: https://arxiv.org/abs/2201.07436

[21] 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.

[22] T. Srinivasan and Y. Bisk, “Worst of both worlds: Biases compound in pre-trained vision-and-language models,” *NAACL Workshop on Gender Bias in Natural Language Processing*, 2021.