2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 14-18, 2024. Abu Dhabi, UAE

ChatMap: A Wearable Platform Based on the Multi-modal Foundation Model to Augment Spatial Cognition for People with Blindness and Low Vision

Yu Hao, Alexey Magay, Hao Huang, Shuaihang Yuan, Congcong Wen and Yi Fang

***Abstract*— Spatial cognition refers to the ability to gain knowledge about their surroundings and utilize this information to identify their location, acquire resources, and navigate their way back to familiar places. People with blindness and low vision (pBLV) face significant challenges with spatial cognition due to the reliance on visual input. Without the full range of visual cues, pBLV individuals often find it difficult to grasp a comprehensive understanding of their environment, leading to obstacles in scene recognition and precise object localization, especially in unfamiliar environments. This limitation extends to their ability to independently detect and avoid potential tripping hazards, making navigation and interaction with their environment more challenging. In this paper, we present a pioneering wearable platform tailored to enhance the spatial cognition of pBLV through the integration of multi-modal foundation model. The proposed platform integrates a wearable camera with audio module and leverages the advanced capabilities of vision language foundation model (i.e., GPT-4 and GPT-4V), for the nuanced processing of visual and textual data. Specifically, we employ vision language models to bridge the gap between visual information and the proprioception of visually impaired users, offering more intelligible guidance by aligning visual data with the natural perception of space and movement. Then we apply prompt engineering to guide the large language model to act as an assistant tailored specifically for pBLV users to produce accurate answers. Another innovation in our model is the incorporation of a chain of thought reasoning process, which enhances the accuracy and interpretability of the model, facilitating the generation of more precise responses to complex user inquiries across diverse environmental contexts. To assess the practical impact of our proposed wearable platform, we carried out a series of real-world experiments across three tasks that are commonly challenging for people with blindness and low vision: risk assessment, object localization, and scene recognition. Additionally, through an ablation study conducted on the VizWiz dataset, we rigorously assess the contribution of each individual module, substantiating the integral role in the model's overall performance.**

I. INTRODUCTION

The prevalence of visual impairment poses a significant global health challenge, impacting over 253 million individuals with a spectrum of social, emotional, and physical difficulties [1]. These challenges extend beyond mere vision loss, contributing to decreased mobility and an increased risk of falls, injuries, and comorbidities, which in turn, exacerbate unemployment rates and quality of life compromises [2]. The World Health Organization (WHO) predicts a continuous rise in these numbers, underscoring the urgent need for effective assistive technologies [3]. Visual impairment, ranging from moderate to severe, hampers an individual's ability to perform visual searches—locating specific targets in cluttered environments—a task that even those with unimpaired vision find challenging [4]. This difficulty is compounded for those with various forms of vision loss, including peripheral, central, or hemi-field vision loss, making it strenuous to navigate or identify objects within their surroundings. Similarly, those with blurred vision, nearsightedness, color-deficient vision, or low-contrast vision face added hurdles in distinguishing objects from their background. Furthermore, the capability to assess potential risks and hazards in one's environment is crucial for ensuring personal safety, a task that demands detailed spatial awareness and understanding [5], [6].

Previous assistive technologies for people with blindness and low vision have made strides by utilizing computer vision for object recognition [7], [8], GPS for navigation [9], and text-to-speech tools for information conveyance [10]. Despite the value they offer, these technologies encounter limitations, particularly in delivering a comprehensive understanding of complex scenes and providing helpful guidance tailored to the specific requirements of visually impaired users. Often, the focus of such tools is on specific functionalities like obstacle detection or route mapping, but they may not provide the nuanced descriptions and contextual information necessary for a more complete and independent interaction with the environment. Moreover, previous systems often restrict visually impaired users to passive reception of interpreted information, lacking interactive capabilities. For example, RFID tag placement framework was proposed for in-building navigation [11], yet such systems may primarily offer one-way communication, limiting user engagement. Additionally, [12] explored blind guidance using mobile computer vision, which, while promising, may not provide sufficient interactivity for users to actively engage with the environment. In contrast to these challenges, our wearable platform addresses these limitations by providing more effective and comprehensive information and enabling active interaction with the environment. By leveraging wearable technology, our platform offers real-time feedback and interaction, thereby enhancing the autonomy and user experience for blind and visually impaired individuals.

In this project, we introduce ChatMap, a wearable platform to augment the spatial cognition of people with blindness and low vision through the integration of proprioception—the in-

[1]Yu Hao, Alexey Magay, Hao Huang, Shuaihang Yuan, Congcong Wen and Yi Fang are with the Embodied AI and Robotics (AIR) Lab, NYU Tandon and NYU Abu Dhabi. Yi Fang is the corresponding author: yfang@nyu.edu

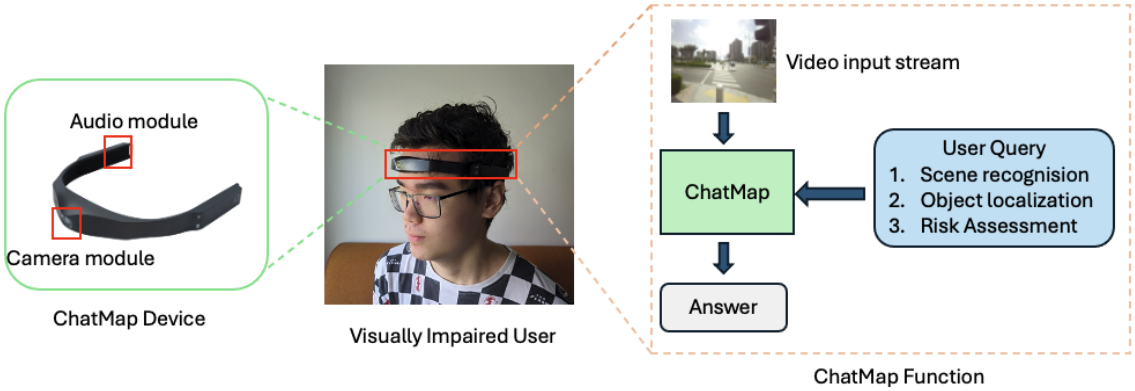

Fig. 1. Overview of ChatMap. (Left) The ChatMap wearable device, featuring a camera module for visual input and an audio module for auditory interaction. (Right) The ChatMap Function, which processes the video input stream alongside user queries to generate answers and guidance for three types of inquiries: scene recognition, object localization, and risk assessment.

nate sense that allows individuals to perceive the position and movement of their body parts without sight, enabling users to intuitively perceive and map their environment relative to their own egocentric coordinate system. As shown on the left side of Figure 1, our platform, equipped with a wearable camera, utilizes a camera module to capture a continuous video stream, providing a real-time depiction of the user's surrounding environment. Additionally, it features an audio module dedicated to receiving user queries and delivering the generated responses audibly to the user. This platform harnesses the capabilities of advanced foundation models to enhance visual perception, encompassing aspects such as scene understanding, object localization, and risk assessment as shown on the right side of Figure 1. By providing users with detailed and comprehensive scene descriptions, along with risk guidance tailored to their inquiries, our approach empowers pBLV individuals with a deeper comprehension of their environment. This facilitates not only the identification and location of objects of interest but also the recognition of potential hazards.

Our wearable platform includes three main components, as illustrated in Figure 2: vision language model, prompt engineering for pBLV, and large language model. Initially, the system employs a vision-language model to extract a detailed proprioceptive description of the user's surrounding environment from an input image. Following this, through the use of prompt engineering, we integrate the scene description with the user's query into a structured prompt, specifically designed to guide the large language model in producing responses that are finely tuned to the needs of pBLV individuals. To ensure the responses are both accurate and comprehensible, we incorporate a chain of thought reasoning process that meticulously analyzes the scene information and user query, delineating the logical steps leading to the final answer. Our experiments demonstrate that our system is able to recognize objects of interest and provide detailed answers to user questions, significantly enhancing the visual understanding of surroundings. Our contributions are summarized as follows: 1) We introduce a voice-controlled wearable platform to augment the spatial cognition of pBLV, capitalizing on the strengths of multimodal foundation models. This advanced platform supports users by providing detailed environmental descriptions, facilitating the recognition of scenes, pinpointing the location of objects, and assessing potential risks. 2) We propose a framework that employs a vision-language model to produce proprioceptive scene descriptions. Additionally, through prompt engineering, we tailor the model to function as a dedicated assistant for pBLV users. The integration of chain of thought reasoning further refines the accuracy and interpretability of responses, ensuring that the answers are both precise and easily understandable for pBLV individuals. 3) We validate our wearable platform's effectiveness through rigorous testing on real-world data and VizWiz dataset. These experiments demonstrate the system's ability to accurately recognize objects and provide accurate descriptions and analyses of the environment, thereby directly addressing the core research problem of enhancing navigation and interaction for pBLV in diverse settings.

II. Related Work

Numerous studies have addressed challenges in assisting blind and visually impaired individuals through various technological approaches [13]. Manduchi and Bagherinia [12] explored the use of mobile computer vision for guiding blind individuals, with a focus on enhancing their mobility and spatial awareness through real-time environmental perception . Their study investigated the feasibility of leveraging mobile devices equipped with vision-based technologies to provide navigational assistance and enhance independent mobility for visually impaired users. Krishna et al. [14] introduces a wearable assistive device for the blind, enabling text-to-audio conversion for enhanced accessibility to printed material. The device, based on Raspberry Pi and equipped with a finger-mounted camera, captures text pointed to by the user and processes words using Optical Character Recognition (OCR). Text-to-Speech (TTS) converter is then used to present that text as audio. The work by Hao et al. [13] proposes using the power of foundation models to augment spatial cognition in individuals with low vision through proprioceptive guidance.

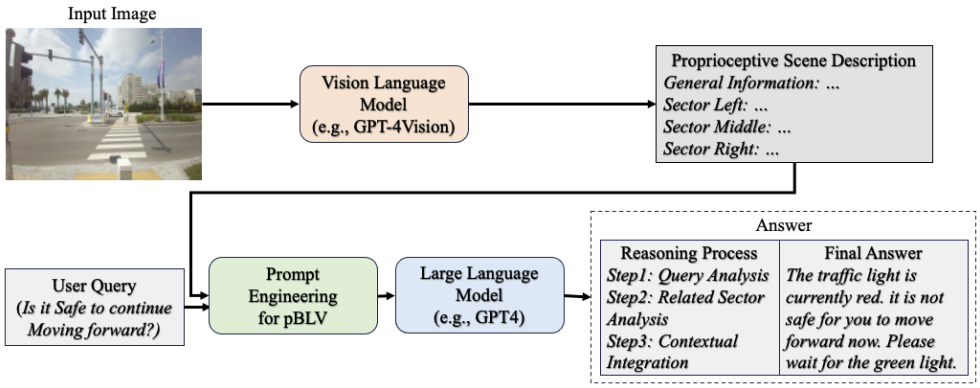

Fig. 2. Flowchart of the ChatMap. Our system comprises three main components: a vision language model for providing a proprioceptive description of the environment, prompt engineering customized for pBLV, and a large language model, enriched by chain of thought reasoning, for response generation.

Advancements in natural language processing have been propelled by the development of language foundation models, designed to comprehend and tackle a diverse array of linguistic tasks using a unified architecture. Models like GPT [15] have set new benchmarks for understanding and generalization capabilities. Each model brings distinct strengths: GPT-2 [16] and BERT [17] laid the groundwork with their deep learning frameworks, while PaLM [18] and GPT-3 [19] expanded the horizons with their vast scale and nuanced language understanding abilities. The latest iteration, GPT-4 [20], builds upon these advancements, offering enhanced language comprehension. These language foundation models, trained on extensive datasets, exhibit robust capabilities for generalization and reasoning across a multitude of tasks. Moreover, their structured approach to information processing enhances their ability to reason through complex tasks, leading to systematic chain-of-thought reasoning [21] and significantly enhancing their proficiency in addressing intricate problems.

III. Method

In this section, we delve into the methodology of proposed wearable platform to augment spatial cognition using the capabilities of multi-modal foundation model. The section is systematically divided into three subsections for clarity and depth. In III-A, we integrate a vision-language model that employs proprioceptive view partition to deliver detailed scene descriptions. III-B discusses the tailored prompt engineering approach that aligns with the specific needs of blind and low vision users. In III-C, we explore the incorporation of chain of thought reasoning, a technique pivotal for refining the accuracy and explicability of the responses provided by the system

A. Proprioceptive Scene Recognition with Vision Language Model

Proprioception, an innate sense crucial for perceiving the position and movement of one's own body parts without visual input, inspires our approach to enhancing spatial cognition for individuals with blindness and low vision [22]. This internal sense is instrumental in navigating and interacting with the surrounding environment, allowing for an intuitive understanding of space relative to one's own body without sight. Our wearable platform, designed to be worn on the head, leverages this fundamental human ability, acknowledging that users naturally move their heads to scan their environment.

Inspired by the concept of proprioception and the instinctive head movements of users, we have engineered our platform to divide the visual field captured by the wearable device into three sectors: left, center, and right. This segmentation mirrors the proprioceptive feedback mechanism, facilitating a more natural and intuitive exploration of surroundings for pBLV. By adapting the device's operation to mimic these head movements, our model not only aligns with how users typically perceive their environment but also significantly enhances their ability to navigate and understand their spatial context.

Once the environment is segmented, the model performs a focused analysis within each sector, akin to how proprioception allows individuals to sense parts of their body in space without visual cues. This is crucial for accurately addressing queries about the location of objects. For instance, if a user inquires, *"Where is the chair?"*, the model detects the chair's presence and responds with its position relative to the user's current orientation, such as *"to your left,"* effectively using proprioceptive-like feedback to guide the user's awareness to the left sector.

The rapid advancement of foundation models, particularly in the realm of foundation models, such as large language model and vision language model, has brought about remarkable capabilities in understanding and reasoning. These models' extensive pretraining on diverse datasets enables them to exhibit a deep comprehension of various contexts and nuances. In leveraging the power of these models, we aim to harness their potential to significantly enhance the functionality of assistive devices for the blind and low-vision community.

Our system incorporates a vision-language foundation model to articulate detailed descriptions of the scenes captured by the smart wearable device. We have chosen GPT-4V [20] for its robust performance to serve as the vision-language foundation model in our system. GPT-4V's nuanced

understanding of visual inputs allows for rich and detailed depictions of the captured scenes, essential for the system to obtain the comprehensive description of their surroundings.

B. Prompt Engineering for pBLV

In our wearable platform designed for individuals with visual impairments, we bring in the prompt engineering [23] to customize interactions specifically for our users. This technique combines the visual insights gathered by the vision language model with the queries posed by users to create prompts that are both contextually relevant and highly informative. Unlike conventional machine learning strategies that rely on extensive datasets for model training, our method focuses on crafting precise prompts to guide the model's responses, bypassing the need for model parameter optimization.

The vision language model generates a detailed description. We include the prompt "*I will provide you a description of what I see now:* {*descriptions*}" to integrate the scene description results into an prompt. Furthermore, user queries are integral to our prompt engineering process, enabling us to tailor the system's responses to the specific informational needs of the user. By embedding these queries within our prompts, we ensure that each interaction is directly aligned with the user's current context and requirements. This method ensures that responses are clear, concise, and tailored to the user's needs, while maintaining a conversational tone that is sensitive and considerate. The model's response is designed to be informative without drawing unnecessary attention to the user's visual impairment, thereby ensuring the communication is both helpful and respectful.

C. Chain of Thought Reasoning

In the section, we explore the integration of a multi-step reasoning approach, inspired by chain of thought reasoning in [21], to enhance the interpretability and accuracy of answers generated by our system. We observed one single-step reasoning processes have often fallen short in providing the level of detail and precision necessary for addressing the complex queries of users, particularly in the context of spatial cognition for individuals with blindness and low vision. To counter this, our approach adopts the chain of thought reasoning, which allows for the sequential construction of a detailed reasoning path, as depicted in reasoning process box in Figure 2. Specifically, we first conduct a query analysis (Step 1) to dissect and understand the user's query, identifying its key elements and objectives. This is followed by a related sector analysis (Step 2), where we examine the relevant sector information within the scene description to ensure the reasoning process incorporates spatial information pertinent to the user's inquiry. Subsequently, through contextual integration (Step 3), we amalgamate the insights obtained from both the query and sector analyses to create a unified and informed basis for the response. By articulating a clear, stepwise reasoning path, our system provides responses that users can easily understand and trust, thereby enhancing the usability and effectiveness of our approach for pBLV individuals.

The initial step in constructing a reasoning chain is to use the user's instruction and the prior reasoning step as inputs to generate the current reasoning output, sequentially building the complete reasoning chain. This step-by-step approach not only facilitates an accurate buildup of reasoning but also enhances the interpretability and reliability of the answer. The formulation of the final answer follows a similar process, applying the reasoning chain, the scene description, and the user query as foundations to sequentially deduce each component of the answer. This structured decomposition into multiple conditional probabilities serves to refine the precision of the output, culminating in a response that is both informative and intelligible.

IV. Experiments

A. Real-World Experiment

Experimental Setting: In our study, we conducted experiments to assess the performance of the our wearable system across various scenarios as shown in Table I. Our system utilizes a head-mounted video camera, specifically the Drift X3 [24], coupled with a smartphone to facilitate seamless interaction between the user and the system. We employed six distinct scenes to evaluate the system's capabilities, posing a total of 24 questions categorized into three domains: risk assessment, object localization, and scene recognition. To establish a baseline for comparison, we manually generated ground truth answers corresponding to each question and scene frame. The distribution of questions across categories was as follows: 7 questions pertained to scene description, 8 focused on object localization, and 9 addressed risk assessment. This methodology allows for both qualitative and quantitative assessments of similarity and quality of responses generated by our system against the manually established ground truth. For quantitative analysis of our system's question answering capabilities we employ common evaluations metrics such as BLEU, ROUGE and METEOR [25]. These metrics evaluate the similarity between system's answers and established ground truth answers on both structural and semantic levels.

Results: Both qualitative and quantitative evaluations of our system's responses demonstrate the performance across three task categories: object localization, scene recognition, risk assessment as shown in Table I and Figure 3. Our system demonstrated best performance on the tasks related to object localization with the scores of 76.6, 69.8, 85.6, and 73.3 for BLEU-1, BLEU-2, ROUGE, and METEOR respectively. This means that most of the answers produced by our system were the exact or very close matches to ground truth, indicating strong assistive capabilities for blind and visually impaired. The metrics for scene recognition are lower (refer to Table I). However, scene recognition task requires more articulation, therefore there are more ways to formulate a correct answer. In other words, in some cases, system's responses scored lower across all metrics due to the variance in language model's text generation, not necessarily due to

	BLEU1	BLEU2	ROUGE	METEOR
Rist Assessment	26.8	11.2	24.8	20.0
Object Locolization	76.6	69.8	85.6	73.3
Scene Recognition	43.1	17.9	58.2	30.0

TABLE I

QUANTITATIVE RESULTS ON THE REAL-WORLD DATASET.

Input image					
Task category	Risk assessment	Object localization	Scene description	Object localization	Scene description
Question	Is green light or red light?	Where is the trash can?	What's in front of me?	Which way is the pharmacy?	Is the palm area crowded?
Model's answer	Red light	The trash can is in the kitchen area on the left side.	In front of you, there is a kitchen workspace with various appliances and utensils, a sink with dishes, and kitchen items scattered on the counter.	The pharmacy is to your left.	No, the palm area is not crowded.

Fig. 3. Qualitative Results of Random Selected Examples on the real dataset.

	VLM	LLM	Prompt Engineering	Proprioceptive Scene Recognition	Chain of Thought Reasoning	BLEU1	BLEU2	ROUGE	CIDER	METEOR
Model 1	✓					25.4	14.5	25.7	19.0	10.4
Model 2	✓	✓				34.7	22.3	28.5	28.7	13.3
Model 3	✓	✓	✓			43.9	26.1	29.7	35.4	15.2
Model 4	✓	✓	✓	✓		46.3	28.3	30.6	38.1	16.4
Model 5	✓	✓	✓	✓	✓	48.0	29.3	32.6	41.2	16.9

TABLE II

QUANTITATIVE RESULTS OF ABLATION STUDY ON THE VIZWIZ DATASET.

wrong reasoning. Inspection of qualitative results indicates strong performance in scene recognition as well.

B. Ablation Study on VizWiz dataset

Experimental Setting: In this section, we conduct experiments to test the effectiveness of each component within our system on the task of Visual Question Answering [26] of VizWiz dataset. The VizWiz dataset [27] is a collection of images taken by blind and visually impaired individuals, specifically designed to evaluate computer vision algorithms aimed at assisting visually impaired individuals. Our evaluation employs metrics such as BLEU, ROUGE-L, METEOR, and CIDEr [28], with a specific focus on BLEU-1 and BLEU-2 due to the typically concise nature of Visual Question Answering responses.

To rigorously test the effectiveness of each component within our system, we explore five distinct model settings. In the first setting, we directly apply a vision language model (i.e., GPT-4) to derive answers from the user query and the current frame as shown in Table II. The second setting progresses by first generating a scene description from the current frame using VLM, then employing a large language model (i.e., GPT-4) to formulate answers based on the textual scene description and user query. The third setting introduces prompt engineering, refining the model to act as an assistant tailored specifically for people with blindness and low vision, aiming to produce answers that are more relevant and supportive. The fourth setting considers human proprioceptive abilities to perceive head movement, integrating a proprioceptive scene description that segments the frame into three distinct sectors—left, center, and right. Finally, in the fifth model setting, we incorporate chain of thought reasoning to further enhance answer accuracy and interpretability by explicitly generating a reasoning chain alongside the final answer. Through these diverse settings, we aim to validate the impact of each system component on improving assistance for pBLV individuals.

Results: Examining the performance outcomes depicted in the table, we observe that using a Vision-Language Model alone yields limited success. While VLM excels in interpreting and describing images, it does not inherently provide the comprehensive reasoning needed to generate precise answers tailored to the specific inquiries of people with blindness and low vision. Consequently, the integration of a Large Language Model with VLM in Model 2 delivers superior performance, highlighting the LLM's contribution to more nu-

anced understanding and response generation. Performance further improves with the integration of prompt engineering tailored for pBLV, confirming its effectiveness in customizing responses. Including proprioceptive scene descriptions that divide the visual input into three sectors—left, center, and right—aligns with natural human spatial cognition, leading to even better system performance. The most significant advancements are observed in the fifth model setting, where the incorporation of chain of thought reasoning markedly enhances answer accuracy and interpretability, as demonstrated by the highest scores across all evaluation metrics. This progression validates the layered approach to system development, with each added component contributing to a more sophisticated and user-centric platform for pBLV individuals.

V. CONCLUSIONS

In this paper, we introduce a novel wearable platform designed to significantly enhance the spatial cognition capabilities of individuals with blindness and low vision. By integrating a multi-modal foundation model, including GPT-4V, the platform offers a nuanced processing of visual and textual data through a wearable camera complemented with audio capabilities. Our system utilizes vision language models for proprioceptive scene recognition, aligned with the natural head movements of users, and employs prompt engineering to tailor the Large Language Model to serve as a personalized assistant for pBLV users. The integration of these elements, alongside the innovative chain of thought reasoning process, ensures that our model delivers not only detailed scene understanding and object localization but also efficient risk assessment. The efficacy of our platform has been validated through real-world experiments and a comprehensive ablation study on the VizWiz dataset, demonstrating significant improvements in the way pBLV users interact with and understand their environment, thereby paving the way for greater independence and safety in their daily lives.

REFERENCES

[1] D. Pascolini and S. P. Mariotti, "Global estimates of visual impairment: 2010," *British Journal of Ophthalmology*, vol. 96, no. 5, pp. 614–618, 2012.

[2] L. Hakobyan, J. Lumsden, D. O'Sullivan, and H. Bartlett, "Mobile assistive technologies for the visually impaired," *Survey of ophthalmology*, vol. 58, no. 6, pp. 513–528, 2013.

[3] W. H. Organization *et al.*, "Visual impairment and blindness fact sheet n 282," *World Health Organization*, 2014.

[4] A. M. Treisman and G. Gelade, "A feature-integration theory of attention," *Cognitive psychology*, vol. 12, no. 1, pp. 97–136, 1980.

[5] H. Fernandes, P. Costa, V. Filipe, H. Paredes, and J. Barroso, "A review of assistive spatial orientation and navigation technologies for the visually impaired," *Universal Access in the Information Society*, vol. 18, no. 1, pp. 155–168, 2019.

[6] Z. Yuan, T. Azzino, Y. Hao, Y. Lyu, H. Pei, A. Boldini, M. Mezzavilla, M. Beheshti, M. Porfiri, T. Hudson *et al.*, "Network-aware 5g edge computing for object detection: Augmenting wearables to" see"more, farther and faster," *arXiv preprint arXiv:2112.13194*, 2021.

[7] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," *arXiv preprint arXiv:1804.02767*, 2018.

[8] Y. Hao, J. Feng, J.-R. Rizzo, Y. Wang, and Y. Fang, "Detect and approach: Close-range navigation support for people with blindness and low vision," in *European Conference on Computer Vision*. Springer, 2022, pp. 607–622.

[9] GPS.gov, "Gps accuracy," *Official U.S. government information about the Global Positioning System (GPS) and related topics*.

[10] N. Li, S. Liu, Y. Liu, S. Zhao, and M. Liu, "Neural speech synthesis with transformer network," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 33, no. 01, 2019, pp. 6706–6713.

[11] W. S. Mooi, T. C. Eng *et al.*, "Efficient rfid tag placement framework for in building navigation system for the blind," in *8th Asia-Pacific Symposium on Information and Telecommunication Technologies*. IEEE, 2010, pp. 1–6.

[12] R. Manduchi, S. Kurniawan, and H. Bagherinia, "Blind guidance using mobile computer vision: A usability study," in *Proceedings of the 12th international ACM SIGACCESS conference on Computers and accessibility*, 2010, pp. 241–242.

[13] Y. Hao, F. Yang, H. Huang, S. Yuan, S. Rangan, J.-R. Rizzo, Y. Wang, and Y. Fang, "A multi-modal foundation model to assist people with blindness and low vision in environmental interaction," *Journal of Imaging*, vol. 10, no. 5, p. 103, 2024.

[14] A. B. Krishna, M. Hari, and A. Sudheer, "Word based text extraction algorithm implementation in wearable assistive device for the blind," in *2019 IEEE International Conference on Intelligent Techniques in Control, Optimization and Signal Processing (INCOS)*. IEEE, 2019, pp. 1–5.

[15] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever *et al.*, "Improving language understanding by generative pre-training," 2018.

[16] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever *et al.*, "Language models are unsupervised multitask learners," *OpenAI blog*, vol. 1, no. 8, p. 9, 2019.

[17] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[18] A. Chowdhery, S. Narang, J. Devlin, M. Bosma, G. Mishra, A. Roberts, P. Barham, H. W. Chung, C. Sutton, S. Gehrmann *et al.*, "Palm: Scaling language modeling with pathways," *arXiv preprint arXiv:2204.02311*, 2022.

[19] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.

[20] OpenAI, "Gpt-4 technical report," 2023.

[21] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou *et al.*, "Chain-of-thought prompting elicits reasoning in large language models," *Advances in Neural Information Processing Systems*, vol. 35, pp. 24 824–24 837, 2022.

[22] W. Gui, B. Li, S. Yuan, J.-R. Rizzo, L. Sharma, C. Feng, A. Tzes, and Y. Fang, "An assistive low-vision platform that augments spatial cognition through proprioceptive guidance: Point-to-tell-and-touch," in *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2019, pp. 3817–3822.

[23] 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.

[24] D. Innovation, "Drift x3." [Online]. Available: https://us.driftinnovation.com/pages/ghost-x3

[25] S. Banerjee and A. Lavie, "Meteor: An automatic metric for mt evaluation with improved correlation with human judgments," in *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, 2005, pp. 65–72.

[26] 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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 3608–3617.

[27] J. P. Bigham, C. Jayant, H. Ji, G. Little, A. Miller, R. C. Miller, R. Miller, A. Tatarowicz, B. White, S. White *et al.*, "Vizwiz: nearly real-time answers to visual questions," in *Proceedings of the 23nd annual ACM symposium on User interface software and technology*, 2010, pp. 333–342.

[28] R. Vedantam, C. Lawrence Zitnick, and D. Parikh, "Cider: Consensus-based image description evaluation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 4566–4575.