



THE JOURNAL ON  
TECHNOLOGY AND  
PERSONS WITH  
DISABILITIES

# iGuide: A Multimodal App for Visually Impaired Museum Visitors

Brandon Vasquez<sup>1</sup>, Rafid Zia<sup>1</sup>, Joshua Henry<sup>1</sup>, Zhigang Zhu<sup>1,2</sup>

City College of New York, CUNY<sup>1</sup>, The Graduate Center, CUNY<sup>2</sup>

[bvasque004@citymail.cuny.edu](mailto:bvasque004@citymail.cuny.edu), [rzia000@citymail.cuny.edu](mailto:rzia000@citymail.cuny.edu),

[jhenry012@citymail.cuny.edu](mailto:jhenry012@citymail.cuny.edu), [zzhu@ccny.cuny.edu](mailto:zzhu@ccny.cuny.edu)

## Abstract

The use of technology in public spaces, such as museums, for visually impaired individuals, is usually done by enhancing the sensory experience for low-vision visitors, primarily with regard to tactility and auditory sensation. However, these experiences do not consider the interactive aspect. Hence, iGuide, a mobile application featuring text-to-speech and speech-to-text with natural language responses from a Large Language Model (LLM), was developed to allow for an interactive multimodal approach to aiding visually impaired users. Additionally, the work considers the use of robotics, specifically the Boston Dynamics SPOT quadruped robot dog in reducing the cognitive load for the visually impaired when paired with the developed system, in terms of visual orientation and navigating museum spaces. These design choices are guided by combining both the state of the art in terms of accessibility software, hardware, and physical implementations as well as the interactivity provided by guides.

## Keywords

Visually Impaired, Large Language Model, App Development, Museums, Text-to-Speech, Speech-to-Text.

## Introduction

According to the International Agency for the Prevention of Blindness, at least 2.2 billion people globally have vision impairment or blindness (Ackland, Resnikoff, and Bourne 71–73). Visual impairment can profoundly impact daily activities, limiting access to information, navigation, and environmental interaction. In settings such as museums, where information is predominantly conveyed through visual means, these challenges are even more pronounced, furthering the need for technologies that can interpret visual data and present it in accessible formats.

Some preexisting technologies are screen readers (American Foundation for the Blind) and TTS systems, such as JAWS, which converts on-screen text into speech, enabling users to access digital content audibly (Freedom Scientific). Mobile applications have also made significant strides; Optical Character Recognition (OCR) tools like KNFB Reader capture text from images and read it aloud, facilitating access to printed materials (National Federation of the Blind). Microsoft's Seeing AI describes people, text, and objects captured by a smartphone camera (Microsoft, Seeing AI), while Be My Eyes (Be My Eyes) connects visually impaired users with sighted volunteers through live video calls for real-time assistance. In the hardware-software space, wearable devices like OrCam MyEye (OrCam) use smart cameras mounted on eyeglasses to read text and recognize faces, providing on-the-go support. While these technologies are valuable, they often are costly, and have limitations in processing complex environments, offering detailed contextual information, and facilitating interactive conversations.

Furthermore, a similar system as ours can be found in (Kuzdeuov et al. 722–727) where the proposed system is a mobile application that leverages the OpenAI GPT to aid the visually impaired and blind. (Park, An, and Park 1541–1544) developed a system that aids visually

impaired users in navigating and understanding museum spaces and art galleries. Brajvasi et al. (Brajvasi and Kirar 1–6) developed a mobile android application that is capable of streaming visual data from a smartphone’s camera into a memory optimized TensorFlow model that allows for deciphering what was captured in the image and providing vocal feedback to the user. In Pujari et al. (Pujari, Madnal, and Premchandran 1–6) the proposed system is a Flutter multi-platform mobile application that broadly aims to improve accessibility using object and currency detection along with an OCR tool for making use of text from images. (Zheng et al. 177–186) proposed a system that leverages a plastic optical fiber (POF) network for high bandwidth between a client-server that makes use of an ai model capable of speech, motion, and image recognition along with LLM’s such as GPT and Vicunna to produce natural responses.

Integrating Large Language Models (LLMs into assistive technologies can enhance interpretative and conversational aspects, providing richer and more interactive experiences for users. Hence, the proposed work *iGuide* involves developing a mobile application that ingests multimodal inputs and uses OpenAI’s GPT-4 LLM Application Program Interface (API) to process these inputs with a TTS-STT pipeline (Fig. 1). By processing inputs through speech, images, or text, the application provides seamless interaction and conversational assistance with its audio playback.

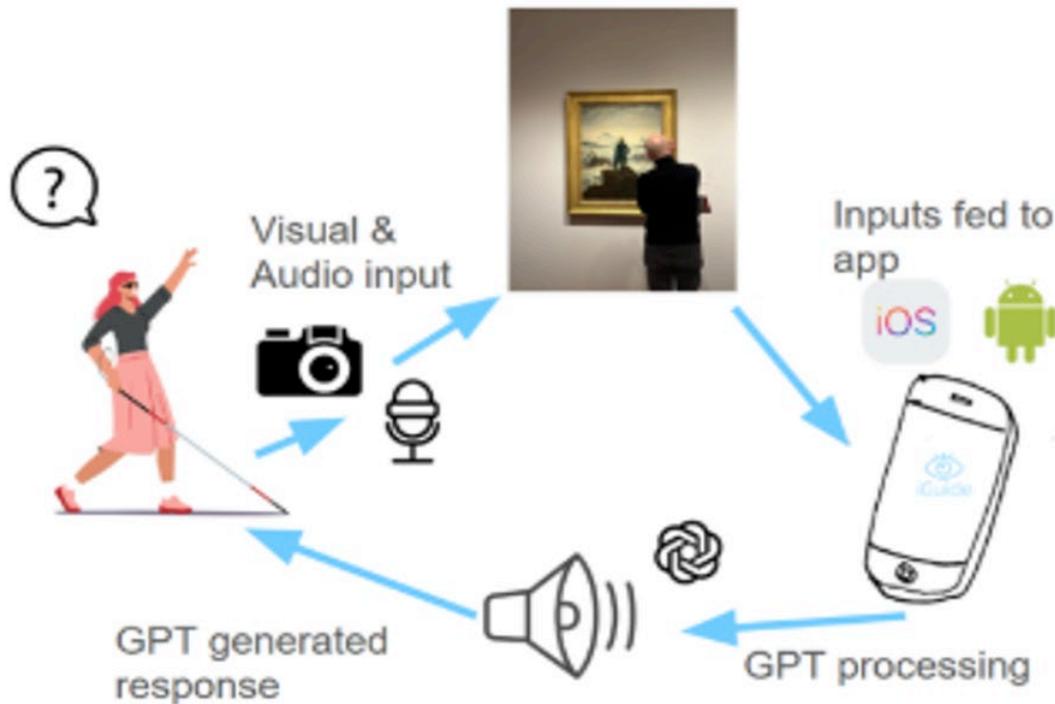


Fig. 1. High-level overview of iGuide end-to-end process. From left to right and back to left (clockwise), the diagram shows the user prompts the system by capturing visual and/or audio input, which then gets fed into the system for processing by the GPT and the TTS-STT.

Supplementary hardware such as the Boston Dynamic's robot dog SPOT (Boston Dynamics) is considered for further improving the conversational efficacy of the system and reducing the cognitive load for users since SPOT is a quadruped robot capable of autonomous movement and is equipped with advanced sensors and cameras and can navigate various terrains, making it suitable as a tour guide or intelligent agent for public spaces such as museums. This will allow SPOT to act similar to a guide, much like the docent from the Smithsonian American Art Museum (Smithsonian American Art Museum).

In summary, by leveraging the reasoning and generative capabilities of LLMs coupled with a TTS-STT pipeline, the aim of the proposed system is to:

- Provide an interactive/conversational experience for users through iGuide using OpenAI's

GPT-4 model for generating responses and multimodal input processing.

- Allow for visually impaired users to interact with museum exhibits through multimodal inputs without the need for additional hardware.
- Explore a potential piece of supplementary hardware, SPOT, that could reduce the cognitive burden on visually impaired individuals when navigating the museum.

## **Discussion**

### *Interface: Portability, Usability and Accessibility*

The interface of the system was developed with a focus on smartphone devices to introduce a seamless system such that no supplementary hardware was necessary, as such the use of the Expo (Expo) framework for the application was considered for two reasons: (1) Ease of shipping to production on multiple platforms, i.e., IOS, Android, and Web. (2) Developing the system with React Native (Meta, React Native) which allowed for a wide plethora of development tools and libraries as well as the systems future maintainability. Furthermore, Expo is a framework used in React Native development workflows which is crucial for the maintainability of the UI for future development.

There were three goals in the development of an effective interface for the visually impaired: accessibility, usability, and portability. The smartphone allows for a wide range of built-in hardware and services that would otherwise need to be sourced by museums for the users as well as would have to be connected in the development phase. This further extended the ease-of-use of the iGuide system that was developed, as it allows users to be able to centralize the use of this application without having to integrate more devices in their outings to museums or the purchase of more hardware.

To maintain accessibility throughout the application, various features from the Expo

framework were used, such as the accessibility label (Expo, Expo Image) in conjunction with built-in screen readers such as VoiceOver (Apple, VoiceOver User Guide). As seen throughout Fig. 2, where the most relevant objects are only a couple buttons on the screen which are large enough to feel around quickly. Inclusively, at each transition of the UI from the homepage there is a voice that announces what is to be done at the landing page e.g. when the voice recording button is clicked on the home page there is a transition screen that describes what the user should prepare to do.

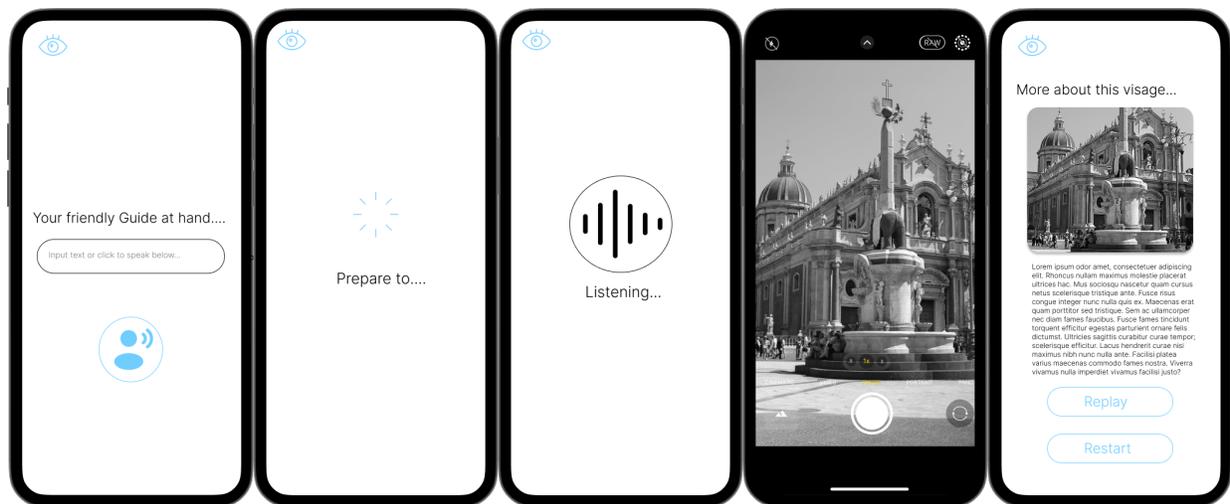


Fig. 2. iGuide mobile UI. From left to right: (1) the landing page, (2) the transition page which directs the user on the upcoming task, (3) the audio recording page, (4) the image capturing page, and (5) the response page, which produces the GPT generated response, the image captured, and audio playback along with next step options to repeat the audio playback or restart the flow to capture information on another exhibit.

### *Software and Hardware*

The application will leverage the GPT-4 LLM provided by OpenAI through their public API, for reasoning user multimodal inputs, including text, and images, and responding accordingly. The backend of the system was developed using the microframework Flask (Pallets)

along with the Python programming language for writing the server-side code. Flask was used to handle the HTTP requests made to the server in a RESTful architecture (Gupta). The TTS-STT pipeline for the backend works in such a way that it first receives a JSON payload from the client usually consisting of text, image, and audio data and from there the audio data—if there was any—would be transcribed using OpenAI’s Whisper model to transcribe the audio. This text data would then be passed on to an Agentic architecture based on the newly recommended LangGraph (LangChain, LangGraph) architecture provided by LangChain (LangChain) which would first decide on whether the GPT-4 model was able to respond to the provided prompt or it needed to do research in order to give an accurate response. If the model did not search online, it would simply respond, with a text output and exit the graph. Otherwise, it would use the Bing API (Microsoft, Bing Web Search API) to search online, and from a sample of 100 sources fetched as a JSON data structure, the GPT model would parse, summarize, and condense that information in its response and exit the graph. Lastly, the text would be converted to audio data using the OpenAI TTS API and sent to the client for playback. Meanwhile, the Retrieval Augmented Generation (RAG) functionalities (Amazon Web Services) were packaged into the TTS-STT pipeline this RAG architecture can be further observed in its logical flow in Fig. 3.

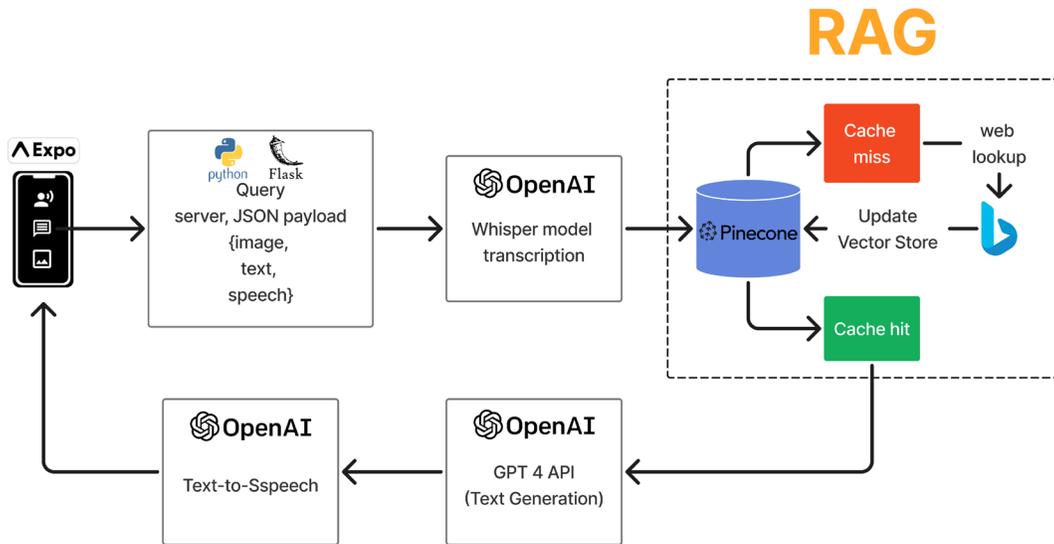


Fig. 3. End-to-end overview of the iGuide system from sending query to client to processing and responding.

The purpose of the RAG architecture, developed with LangChain (LangChain, The Platform for Reliable Agents) was to store information that was retrieved online or more in-depth information already held by a museum, allowing more relevant responses to be generated. This was to make the application more effective in terms of GPT-4's reply quality by using a vector database (Schwaber-Cohen; Pinecone), with its direct integration with Amazon Web Services (AWS Marketplace) for deployment. For the TTS-STT pipeline, the Whisper model provided by OpenAI was used along with their transcription API. Lastly, the system was prepped with the understanding that the GPT-4 LLM being used was to serve as an assistant at a museum.

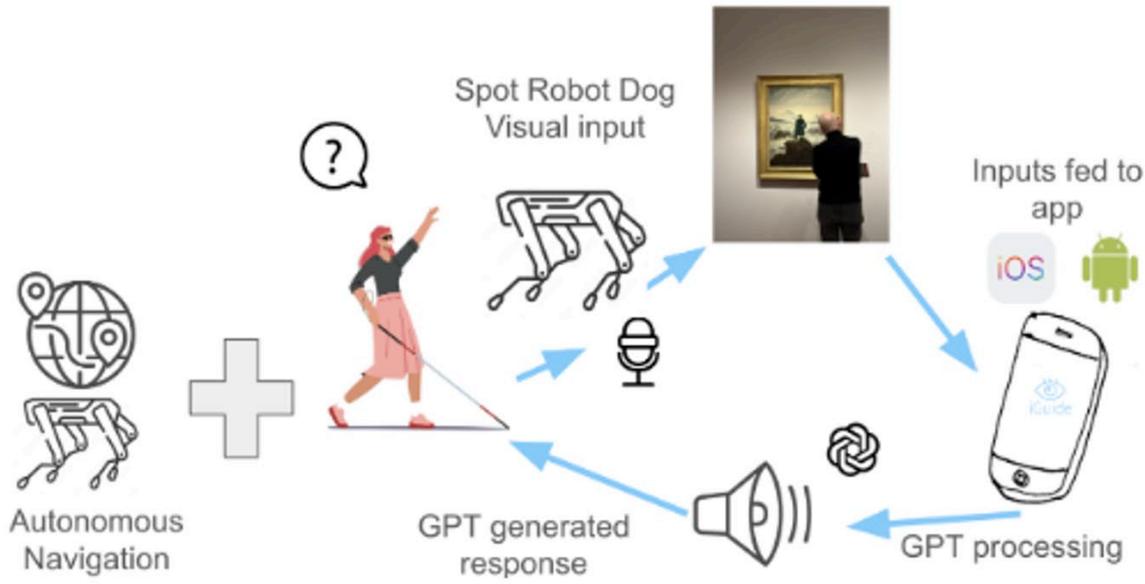


Fig. 4. Overview of the robot SPOT integrated into the system.

Along with the mobile interface, the application can, while not necessary, be used with supplementary hardware such as the SPOT quadruped robot dog from Boston dynamics. The concept of integrating SPOT into the use of the iGuide application stems from the versatility of SPOT which includes autonomous navigation and a built-in camera for streaming visual feed as well as capturing images. Additionally, the robot dog can be vertically scaled to introduce panoramic cameras, or more high-definition cameras for better image analysis by GPT-4 (Fig 4).

## Conclusion

### *Experiment: General*

For testing the developed iGuide system, a large museum was chosen with two reasons in mind. First, a large museum would afford a diverse set of exhibits which the system could ingest and analyze. Second, using the app in such an environment allowed for testing the way in which the system would respond to noise, primarily other people, when capturing images. Hence, the museum chosen for the experiment was the Metropolitan Museum in New York City since it afforded both, a diverse set of exhibits and large crowds. The exhibits that were used for the

experiment consisted of ancient stonework, and immersive interiors from the 18<sup>th</sup> century. Additionally, diverse prompts were used in the experiment to further understand how vague a prompt could be such that the system would be able to answer. The video of the experiment being carried out is available online at: <https://youtu.be/XUv9-aDUmaI>.

To give a brief overview of the first exhibit from the experiment video, the app when booted springs to life and vocalizes some cached audio using the same voice used throughout the TTS process, these samples were collected en masse through a separate piece of code that was made when first developing the system which uses the speech generation model from OpenAI to create a large batch of wav files. The user then selects to input a vocal message to the system, and a transition screen announces to the user to “Prepare to record your voice!” Subsequently the user records his voice with the prompt “tell me about this exhibit?” The app then preps the user to take an image by once again announcing the following task and the user can then feel around the UI for their desired option such as capturing, uploading, or opting out of capturing an image. Finally, the server receives the payload of the image and audio which is processed and sent to the client where the audio is played, and the system vocalizes its analysis. This experiment showcased the accessibility considered for the application in its vocal transitions, and intuitive flow, as well as showcasing the capability of the responses provided by the system which aim to be describe exhibits with senses besides sight.

#### *Experiment: Specifics*

One major focus of our test was to determine whether the model could provide detailed descriptions of elements such as exhibit materials, textures, or painting techniques, including brush strokes or layering. Therefore, the system was prompted with a batch of six images that each was unique in their own fashion and were captured in one of the authors’ rooms and in the

Metropolitan Museum in New York City. As seen in Table 1, the responses were able to successfully point out the material and even brush strokes of certain piece when prompts were provided. Additionally, the average time for the systems latency can be seen as roughly around 14 seconds responses which is admittedly not the best when users may be in an area with poor network or extremely complex queries. Nevertheless, this proved the system was capable of drawing in the user with a sense of immersion through descriptive words but reiterates the point and need for improving the RESTful architecture to a more stream-based one such as Websockets (Websockets) or WebRTC (WebRTC) so real-time interaction can be achieved.

Table 1. Column 1: Image used; Column 2: Response latency T (in seconds); Column 3: Prompt to GPT; Column 4: GPT Response.

Image	T(s)	Prompt	Response
	18	What are the brush strokes on the piece like and what are the facial expressions and emotions of the people in the piece?	The painting features a dark atmosphere with rich, deep colors. The brush strokes appear to be somewhat textured, giving a sense of depth and movement. The figures exhibit various facial expressions: the man seems contemplative, while the women show focused engagement. Their emotions reflect a serious moment, possibly involving discussion or decision-making. The surrounding details add to an intimate yet tense dynamic among the subjects.
	14	What does the coloration and aging of the various blades look like and what does it make one feel?	The image features a collection of historical blades displayed within a case each sword has distinct textures and surfaces. Some are smooth while others display intricate engravings. The metal showcase is various colors with signs of rust and tarnish, indicating age, and wear. The overall presentation evokes a sense of respect and curiosity, inviting you to imagine the stories behind these beautifully crafted weapons. Would you like to know more about the materials or textures?

Image	T(s)	Prompt	Response
	13	How tall is the armor relative to me and what impact does it give?	The armor displayed appears to be encased in glass, suggesting it's quite valuable. You can likely feel a smooth texture from the glass and a contrast with the armor's rugged surface. The armor's height isn't specified, but it looks reasonably tall, possibly reaching to the waist or higher of an average person. This type of armor typically provided significant protection on the battlefield, embodying both strength and craftsmanship. The mesmerizing detail suggests it might have intricate patterns or layers.
	11	Is this piece standing alone or around other furniture and what are the materials?	The piece, a tall clock, stands alone against a wall with no visible surrounding furniture. It features a combination of black and gold materials, likely wood with gilded accents. The surface has ornate carvings, adding a refined texture, and the clock face is prominently displayed. The overall design is elegant, and there's a small information plaque next to it detailing more about the piece. The lighting seems warm, highlighting its craftsmanship.
	15	What is it made of, and how does that material look?	The image features two armored horses and their riders, showcasing suits of armor that are typically made of metal, often steel. The armor has a shiny, polished surface, reflecting light and giving a smooth texture. The wooden plinth beneath them is solid and has a warm, natural grain, contrasting with the cold hardness of the metal armor. Overall, the scene is well-lit, emphasizing the intricate details of the armor and the proud stance of the horses.
	10	In which direction is the beheaded medusa looking at and what is the posture of the man and the mood it conveys emphasizes foot and hand positions and how the light in the gallery helps evoke more emotions	The statue depicts a muscular man holding a sword in one hand and the beheaded Medusa's head in the other. He gazes slightly upward, suggesting triumph or victory. His stance is strong and confident; his feet are apart, grounding him, while his hand grips Medusa firmly. The light in the gallery highlights the smooth textures of the marble, creating depth and emphasizing the man's muscular form and the intricate details of Medusa's face, evoking a sense of awe and tension in the scene.

### User Interviews

User interviews were conducted with four blind individuals at the Visions Center in Manhattan, New York. The data in Table 2 summarizes participant demographics and key feedback points. From the pool of participants, two occasionally visited museums (defined as a

few times a year), one never visited since childhood, and the other rarely visited. The group consisted of three individuals who are completely blind and one who is visually impaired with some vision. All interviews were audio-recorded and transcribed verbatim. The resulting qualitative data underwent thematic analysis, a systematic process for identifying, analyzing, and reporting themes within the data. This involved filling out a Google form for each of the participants with their corresponding transcribed responses.

User interviews yielded rich qualitative data and confirmed the positive impact of the iGuide app prototype. Quantitatively, three out of four participants (75%) expressed dissatisfaction with current museum accessibility (defined as solely Human Guides). A key theme that emerged was the strong preference for independent exploration as well as descriptive explanations of exhibits. After hearing the video of the iGuide prototype demo, a user rebounded, *“If I can go into a museum, have that describe me like that, that’s awesome!”* Participants who occasionally visit museums were particularly responsive to the apps' storytelling and descriptive responses on exhibits.

It was found that the accessibility barriers users face consisted primarily with difficulty locating, and identifying exhibits in addition to an inhibition of their independence. The findings suggest that the iGuide application addresses accessibility barriers identified by participants, primarily the reliance on human guides which can limit independent exploration of the exhibits. The positive feedback on the descriptive audio portion indicates the prototype effectively meets a key user need that aligns with WCAG 2.1 (*Web Content Accessibility Guidelines 2.1*). It was also found that all of the participants used the *Be My Eyes* application (Be My Eyes) to some extent for image capturing with the help from sighted volunteers through live video calls, which bridged an important understanding of what visually impaired users would use for capturing

images and how the image capturing in the system could be made more effective. Lastly, while these findings provide valuable insights into the potential of iGuide, several limitations must be acknowledged when interpreting the results. The primary drawback is the small sample size of participants. Hence further large-scaled studies must be conducted for further insights.

### *Summary and Future Work*

As our unique contributions to the community, iGuide introduces a mobile-only approach to accessible museum interaction by integrating computer vision, text-to-speech, speech-to-text, and large language model tech into a single conversational platform. Unlike prior systems which have been explored which rely on specialized hardware or pre-scripted audio guides, iGuide enables natural dialogue between the user and the system using only a smartphone. This emphasis on minimizing device burden, increasing independence, and conversational approach help in setting iGuide apart as user interviews have even noted on the refreshing approach to the exhibit descriptions. Furthermore, the framework is very versatile and, with minimal changes, can be adapted to other visually intensive public spaces.

The developed iGuide application was tested in a real-world environment in the Metropolitan Museum of New York, which resulted in the understanding of the shortcomings and strengths of the system. The app was able to clearly communicate on intricate and diverse exhibits ranging from sculptures, jewelry, excavated stonework, to paintings. The shortcomings of the app were primarily founded in latency issues which resulted in a lack of conversational feeling with the application. In addition, some future features that would further enhance the usability of the system would be a sight-mode, where the app uses an image detection model to identify surroundings from a live feed and call them out with a TTS model, and once the desired item is within the frame a user can choose to send that image to the system for analysis.

*Acknowledgement*

This article was developed under an appointment to the DHS Summer Research Team Program for Minority Serving Institutions, administered for the U.S. Department of Homeland Security (DHS) by the Oak Ridge Institute for Science and Education (ORISE), managed by Oak Ridge Associated Universities (ORAU) under DOE contract number DE-SC0014664. This document has not been formally reviewed by DHS. The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of DHS, DOE, or ORAU/ORISE. This work has also been supported by the National Science Foundation (NSF) through Award #2131186 (CISE-MSI), and Disability and Accessibility Services at the CUNY Central Office of Student Affairs. The collaboration with and support by Rutgers, Lighthouse Guild, Jose Ozuna and all Visions participants, CCNY CEN Course Innovation Program, and CCNY Zahn Innovation Center are greatly appreciated.

## Works Cited

- Ackland, Peter, Serge Resnikoff, and Rupert Bourne. "World Blindness and Visual Impairment: Despite Many Successes, the Problem Is Growing." *Community Eye Health*, vol. 31, no. 102, 2018, pp. 71–73.
- Amazon Web Services. "Pinecone Vector Database – Pay as You Go Pricing." *AWS Marketplace*, <https://aws.amazon.com/marketplace/pp/prodview-xhgyscinlz4jk>. Accessed 17 May 2025.
- Amazon Web Services. "What Is Retrieval-Augmented Generation (RAG)?" *Amazon Web Services*, <https://aws.amazon.com/what-is/retrieval-augmented-generation/>.
- American Foundation for the Blind. "Screen Readers." *AFB*, <https://www.afb.org/blindness-and-low-vision/using-technology/assistive-technology-products/screen-readers>.
- Apple. "VoiceOver User Guide." *Apple Support*, <https://support.apple.com/guide/voiceover/welcome/mac>.
- Be My Eyes. "Let's See the World Together." *Be My Eyes*, <https://www.bemyeyes.com/>.
- Boston Dynamics. "Spot: The Agile Mobile Robot." *Boston Dynamics*, <https://bostondynamics.com/products/spot/>.
- Brajvasi, A., and B. S. Kirar. "Android Based Real-Time Picture-to-Audio Converter for the Visually Impaired." *IEEE International Students' Conference on Electrical, Electronics and Computer Science (SCEECS)*, Bhopal, India, 2024, pp. 1–6. [doi:10.1109/SCEECS61402.2024.10482202](https://doi.org/10.1109/SCEECS61402.2024.10482202).
- Expo. "Create Amazing Apps That Run Everywhere." *Expo Documentation*, <https://docs.expo.dev/>.

Expo. “Expo Image.” *Expo SDK Documentation*,

<https://docs.expo.dev/versions/latest/sdk/image/#components>.

Freedom Scientific. “JAWS.” *Freedom Scientific*,

<https://www.freedomscientific.com/products/software/jaws/>.

Gupta, L. “What Is REST?” *REST API Tutorial*, 2025, <https://restfulapi.net/>.

Kuzdeuov, A., O. Mukayev, S. Nurgaliyev, A. Kunbolsyn, and H. A. Varol. “ChatGPT for Visually Impaired and Blind.” *International Conference on Artificial Intelligence in Information and Communication (ICAIIIC)*, Osaka, Japan, 2024, pp. 722–727.

[doi:10.1109/ICAIIIC60209.2024.10463430](https://doi.org/10.1109/ICAIIIC60209.2024.10463430).

LangChain. “Balance Agent Control with Agency.” *LangGraph*,

<https://www.langchain.com/langgraph>.

LangChain. “The Platform for Reliable Agents.” *LangChain*, <https://www.langchain.com/>. Meta.

“React Native.” *Meta Open Source*, <https://reactnative.dev/>.

Microsoft. “Bing Web Search API.” *Microsoft Bing APIs*, <https://www.microsoft.com/en-us/bing/apis/bing-web-search-api>.

Microsoft. “Seeing AI: Talking Camera for the Blind.” *Microsoft AI*,

<https://www.microsoft.com/en-us/ai/seeing-ai>.

National Federation of the Blind. “KNFB Reader.” *National Federation of the Blind*,

<https://nfb.org/programs-services/knfb-reader>.

OrCam. “Independence Wherever You Go.” *OrCam MyEye*, <https://www.orcam.com/en/myeye/>.

Accessed 17 May 2025.

Pallets. “Flask.” *Pallets Projects*, <https://flask.palletsprojects.com/en/stable/>.

Park, W., H. An, and S. Park. “Smart Docent System for Visually Impaired People with Mobile-Based Artificial Intelligence of Things.” *TENCON 2024 – IEEE Region 10 Conference*, Singapore, 2024, pp. 1541–1544. doi:10.1109/TENCON61640.2024.10902945.

Pinecone. “The Vector Database for Scale in Production.” *Pinecone*, <https://www.pinecone.io/>.

Pujari, V., K. Madnal, and D. Premchandran. “Mobile App for Enhancing Accessibility Among the Visually Impaired.” *2nd DMIHER International Conference on Artificial Intelligence in Healthcare, Education and Industry (IDICAIEI)*, Wardha, India, 2024, pp. 1–6. [doi:10.1109/IDICAIEI61867.2024.10842749](https://doi.org/10.1109/IDICAIEI61867.2024.10842749).

Schwaber-Cohen, Rinat. “What Is a Vector Database & How Does It Work? Use Cases + Examples.” *Pinecone*, <https://www.pinecone.io/learn/vector-database/>.

Smithsonian American Art Museum. “America InSight: Verbal Description Tours.” *Smithsonian American Art Museum*, <https://americanart.si.edu/education/adult/verbal-description-tours>.

WebRTC. “WebRTC.” *WebRTC*, [webrtc.org](https://webrtc.org).

“Websockets.” *Websockets*, [websockets.readthedocs.io/en/stable](https://websockets.readthedocs.io/en/stable).

Web Content Accessibility Guidelines (WCAG) 2.1. <https://www.w3.org/TR/WCAG21/#audio-description-or-media-alternative-prerecorded>

Zheng, J., et al. “Transforming Daily Tasks for Visually Impaired Seniors: Leveraging an OpenAI Model–Enhanced Multimodal Dialogue System with Speech Recognition in POF Smart Homes.” *SoutheastCon 2024*, Atlanta, GA, USA, 2024, pp. 177–186. doi:10.1109/SoutheastCon52093.2024.10500175.