CTU Journal of Innovation and Sustainable Development
ISSN 2588-1418 | e-ISSN 2815-6412

DOI:10.22144/ctujoisd.2025.057

Naviblind: A multimodal AI assistant for visually impaired users to identify product information from images and speech

Minh-Quan Tran[1*], Duy Truong[2], Duy-Tan Pham[3], Minh-Anh Nguyen[4], Duc-Tung Le[5], Di-Hao Le[6], and Quang-Huy Duong[4]

[1]*Faculty of Computer Science, University of Information Technology, Vietnam National University, Viet Nam*

[2]*Faculty of Computer Science and Engineering, Ho Chi Minh City University of Technology, Vietnam National University, Viet Nam*

[3]*Department of Electrical and Electronics Engineering, Ho Chi Minh University of Technology, Viet Nam*

[4]*Faculty of Information Technology, FPT University, Viet Nam*

[5]*Faculty of Information and Communication Technology, Hanoi University of Science and Technology, Viet Nam*

[6]*College of Information and Communication Technology, Can Tho University, Viet Nam*

**Corresponding author (22521191@gm.uit.edu.vn)*

Article info.

Received 30 Jun 2025
Revised 18 Aug 2025
Accepted 27 Sep 2025

Keywords

Human-centered design, multimodal assistive AI, product accessibility, text-to-speech, Vietnamese speech recognition, vision-language models

ABSTRACT

People with visual impairments often face significant challenges in identifying and accessing product information in their daily lives, particularly when visual cues such as packaging details, labels, or expiration dates are inaccessible. In this paper, we present NaviBlind, a multimodal AI-powered assistive system designed to help visually impaired individuals understand key product details through natural interactions. Our system combines image understanding using Gemini Flash vision models with Vietnamese speech recognition powered by PhoWhisper for extracting information needs directly from user voice commands. By uploading an image of the product and speaking what kind of information is needed, such as name, color, type, or expiry date, the system analyzes the image and returns a concise, structured textual description, which is then converted into Vietnamese speech. To ensure reliability, we incorporate mechanisms to detect uncertain or hallucinated outputs from the vision model, especially in cases of low-quality images. The system is deployed as a user-friendly web application, enabling real-time accessibility for users with limited visual capabilities. Experimental evaluation demonstrates the potential of NaviBlind in promoting autonomy and independence for the visually impaired in everyday shopping and product recognition tasks.

1. INTRODUCTION

Individuals with visual impairments encounter significant challenges in independently accessing information about physical objects, particularly commercial products in daily environments. While advancements in assistive technologies such as screen readers and object recognition tools have contributed to improving accessibility, a substantial

gap remains in enabling non-visual interaction with detailed product attributes such as product name, type, colour, and expiration date, etc. Especially in real- world scenarios like shopping or home use.

Traditional solutions for product recognition often rely on barcodes, QR codes, or tactile markers. However, these methods are limited in scope and practicality. For instance, barcodes may be absent or damaged, and tactile systems require manual pre-labeling, which is not scalable for diverse and dynamic products. Furthermore, many existing tools do not allow users to specify what information they wish to extract, leading to either excessive or insufficient details, which hampers usability.

In recent years, the emergence of multimodal learning and the integration of vision and language understanding have opened new avenues for developing accessible AI systems. The fusion of visual understanding models with natural language interfaces allows users to interact with systems in a more intuitive and personalized way. Particularly, advances in vision-language models (e.g., Gemini, ChatGPT) and automatic speech recognition (ASR) models (e.g., PhoWhisper) have enabled systems to interpret visual scenes and respond to human voice queries with increasing precision and fluency.

In this work, we introduce NaviBlind, a multimodal AI assistant designed to empower visually impaired users by enabling them to extract meaningful and customized product information through image and speech inputs. The core contributions of this work are as follows:

– *Multimodal input pipeline:* Users interact with the system by uploading an image of a product and speaking a query in Vietnamese describing the type of information they need (e.g., "Tên sản phẩm và hạn sử dụng"). The speech input is transcribed using PhoWhisper-medium, a Vietnamese fine-tuned ASR model based on Whisper.

– *Contextual visual understanding:* The transcribed user query is used to construct a structured prompt guiding the Gemini 2.0 Flash vision-language model to extract relevant product information. This approach ensures that the generated response is tailored to the user's actual needs, rather than providing generic or irrelevant descriptions.

– *Accessibility-focused output:* The generated description is not only shown on the screen but also converted into Vietnamese audio output using a TTS engine, ensuring usability for blind or low-vision users.

– *Hallucination Detection and Robustness Handling:* To mitigate the risk of hallucinated responses - a known limitation of large generative models, we implement mechanisms to detect and respond appropriately to low-confidence outputs, such as blurry or incomplete images, thus enhancing reliability and safety.

The system was evaluated on a set of real-world product images and voice queries collected in Vietnamese, demonstrating its effectiveness in supporting non-visual product recognition tasks.

2. RELATED WORKS

We categorize related works into three major components of the NaviBlind system: (1) AI-powered assistive technologies for visually impaired, (2) vision-language models (VLMs) for multimodal understanding, and (3) text-to-speech (TTS) systems for audio feedback. In addition, we briefly review AI-powered assistive technologies for visually impaired users.

2.1. Assistive AI systems for the visually impaired

Several AI-powered applications have been developed to assist visually impaired individuals in daily life. Seeing AI by Microsoft Garage (2024) offers scene and text recognition via smartphone camera, while Be My Eyes (2025) connects blind users with sighted volunteers for visual assistance through live video. Lookout by Google (2018) helps users identify objects, read text, and navigate environments. However, most of these systems are limited to English and often rely on cloud-based services or human assistance (Google, 2018). To the best of our knowledge, no previous work has proposed a fully automated multimodal assistant designed specifically for Vietnamese speakers. NaviBlind is among the first systems to bridge this gap.

2.2. Text-to-Speech (TTS)

For voice input, text-to-speech plays a critical role in enabling natural interaction. Whisper by OpenAI (2022) is a state-of-the-art open-source ASR model supporting over 90 languages, including Vietnamese. Compared to cloud-based services such as Google Speech-to-Text (2018), Whisper can be deployed locally, providing more privacy and robustness in offline or low-connectivity environments (Google Cloud, 2018). In NaviBlind,

we use PhoWhisper, a fine-tuned Vietnamese ASR model based on Whisper, to transcribe user queries accurately and seamlessly as the first step of the processing pipeline.

2.3. Vision-Language Models (VLMs)

Recent advances in multimodal models have enabled systems to understand and generate language grounded in visual inputs. Models such as Gemini (Team et al., 2023), BLIP-2 (Li et al., 2023), MiniGPT-4 (Zhu et al., 2023), and Vintern-1B (Doan et al., 2024) demonstrate strong performance in visual question answering, image captioning, and document understanding. Among these, Vintern-1B is one of the few vision-language models specifically pre-trained for Vietnamese, enabling more accurate and culturally relevant responses in Vietnamese language settings. Similarly, LaVy (Tran & Thanh, 2024) is designed for multimodal understanding in Vietnamese and shows promising results in image-text retrieval and reasoning. In NaviBlind, we leverage a VLM module capable of processing Vietnamese input to extract key product information from images and generate natural-language responses. Unlike prior systems that are limited to English or only generate captions, NaviBlind's VLM module is optimized for Vietnamese multimodal interaction and domain-specific tasks such as product identification and question answering.

3. METHODOLOGY

In this section, we present the architecture and implementation details of the NaviBlind system, a multimodal assistive framework that enables visually impaired users to recognize and access product information through images and spoken queries. The pipeline integrates four main components: Audio Input, Image Input, Prompt Engineering, and Audio Output, as shown in Figure 1. The system takes an image of a product and a natural voice query in Vietnamese as inputs, then generates a concise and informative description in both textual and spoken form.

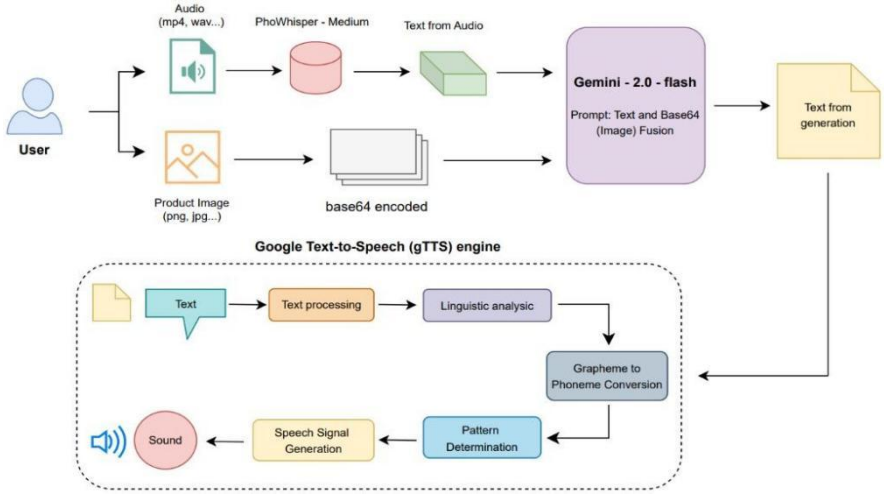

Figure 1. Multimodal pipeline for product recognition and audio feedback in NaviBlind

3.1. Speech-to-text with PhoWhisper

To understand and adapt to the unique information needs of visually impaired users, the first component in our pipeline is a robust Automatic Speech Recognition (ASR) module. In this work, we adopt PhoWhisper-medium, a Vietnamese-tailored ASR model based on OpenAI's Whisper architecture (Le et al., 2024). While Whisper has already shown impressive multilingual capabilities (Radford et al., 2023), PhoWhisper further enhances its performance specifically for Vietnamese by fine-tuning on large-scale, native Vietnamese speech datasets.

PhoWhisper-medium leverages the same encoder-decoder Transformer backbone as the original Whisper models but benefits from additional domain-specific training focused on Vietnamese, making it more accurate and responsive to the nuances of local pronunciation, tonal variation, and informal speaking styles. Among its variants, the medium version offers an optimal trade-off between recognition quality and inference efficiency, making it particularly suited for deployment in real-time,

low-resource environments such as assistive devices for the visually impaired.

Recent evaluations have confirmed that PhoWhisper-medium achieves state-of-the-art performance on Vietnamese ASR benchmarks, consistently outperforming both general-purpose multilingual models and proprietary Vietnamese ASR systems, especially in noisy, spontaneous, or conversational conditions. These characteristics are essential for real-world accessibility applications, where audio input can vary significantly in clarity, device quality, and background noise.

In our system, users either upload or record a short voice clip to specify the product information they want to retrieve (e.g., "tên sản phẩm và hạn sử dụng" - product name and expiration date). To ensure maximum ASR accuracy and consistency, all uploaded audio files are converted into mono-channel WAV format with a 16kHz sampling rate using the pydub library. This normalization step reduces the variability caused by different recording devices and environments.

Once preprocessed, the audio is transcribed into Vietnamese free-form text using PhoWhisper-medium. This text serves as the dynamic query guiding the subsequent vision-language module. By allowing users to speak naturally in their native language without predefined commands or rigid templates, this module significantly enhances usability, especially for non-technical users or the elderly, and reduces reliance on physical or textual interfaces.

3.2. Visual understanding with Gemini

After capturing the user's intent through the audio module, the system proceeds to the second core component: extracting product-related visual information from the uploaded image. To achieve this, we employ Gemini 2.0 Flash, the multimodal foundation model developed by Google DeepMind. Gemini represents a new generation of Vision-Language Models (VLMs) capable of understanding and generating content conditioned on both image and text inputs (Team et al., 2024).

The product image is first converted into a base64-encoded string and passed to Gemini along with the user's transcribed query from the previous module. The prompt is designed to reflect natural Vietnamese expressions. Gemini is then tasked with interpreting the image and producing a concise yet semantically rich description of the relevant product details, such as brand name, color, expiration date, and usage instructions.

The selection of Gemini 2.0 Flash as the core visual understanding module is driven by its advanced multimodal reasoning capabilities, fast response time, and strong support for the Vietnamese language. As a vision-language foundation model trained on large-scale paired image-text data, Gemini demonstrates remarkable performance in understanding complex semantic relationships between visual inputs and natural language prompts. This makes it particularly effective for handling diverse real-world product images where label designs may vary significantly in layout, language, or visual quality. In addition to its reasoning power, Gemini offers low-latency inference through Google's optimized cloud infrastructure, enabling near real-time interaction essential for assistive applications. Importantly, Gemini supports Vietnamese fluently, allowing it to generate coherent and grammatically correct answers that align well with the user's native language. These characteristics make Gemini a suitable choice for our system, offering both accessibility and performance in a language-specific, assistive context.

3.3. Prompt generation with user-specific context

To bridge the gap between user intent and machine interpretation, especially in the context of assistive technology for visually impaired users, we adopt a structured yet flexible prompt engineering strategy. Instead of requiring users to construct complex or explicitly formatted queries, the system leverages natural language input extracted from voice. Users are only expected to describe in their own words the type of information they wish to retrieve from a product image such as its name, color, or expiration date. This simplicity is crucial for accessibility, allowing users with little to no technical background to interact with the system intuitively.

Internally, the system dynamically embeds the user's spoken request into a predefined instruction scaffold designed to guide the Gemini model's behavior. This scaffold ensures that responses are generated in Vietnamese, using complete sentences that include a clear subject and predicate, thereby improving both linguistic clarity and compatibility with downstream speech synthesis. Additionally, the system constrains the response length through a token limit to avoid excessive or irrelevant detail.

An important design consideration is the inclusion of fallback behavior when visual information is insufficient or unclear. The structured prompt guides the model to explicitly indicate, rather than speculating or generating fabricated answers, when requested details cannot be determined from the image such as in cases of blurring, poor lighting, or occlusion. This enhances the system's trustworthiness and ensures that the user is not misled by overconfident hallucinations.

By offloading the complexity of prompt formulation to the system and requiring only minimal natural input from the user, the solution significantly lowers the barrier to access. This design choice not only enables scalable use across diverse product types but also reinforces the system's alignment with its core assistive mission: to empower visually impaired individuals with accurate and meaningful information in an intuitive and human-centered manner.

3.4. Text-to-speech with gTTS

To accommodate users with complete or severe visual impairments, the final stage of the system involves converting textual product descriptions into spoken language. For this purpose, we employ the Google Text-to-Speech (gTTS) (2025) engine, an open- source API that provides fast and natural-sounding speech synthesis in multiple languages, including Vietnamese (Pndurette, 2025). The decision to adopt gTTS is motivated by its combination of simplicity, accessibility, and quality, which makes it particularly suitable for lightweight and cost-effective assistive applications.

gTTS is cloud-based and requires minimal computational overhead on the client device, which aligns well with the constraints of our deployment scenario, particularly in low-resource settings such as charitable or community-focused environments where users may access the system via low-power hardware. Moreover, gTTS offers strong phonetic and prosodic modeling in Vietnamese, producing speech that is not only intelligible but also comfortable for native listeners to follow.

Once the text output is generated by the Gemini model, the system invokes gTTS to synthesize speech and saves it in MP3 format. This audio is immediately made available through the user interface for playback, completing the multimodal interaction loop. The use of MP3 ensures broad compatibility across browsers and devices, while reducing file size and streaming latency. gTTS was chosen not only for its technical efficiency and quality but also for its alignment with the project's mission: to deliver a fully audio-driven interaction pipeline that minimizes user effort while maximizing informational clarity and accessibility.

4. EXPERIMENTS

To evaluate the effectiveness of our proposed NaviBlind system in real-world Vietnamese Visual Question Answering (VQA) scenarios, we conduct a series of controlled experiments focused on product-related image understanding. Our goal is to verify whether state-of-the-art multimodal large language models (MLLMs), when guided by natural language prompts derived from speech, can correctly extract structured and context-aware product information for visually impaired users.

We adopt a single high-quality dataset, curated specifically for this task, and compare some open and API-based VQA models. The choice of models is informed by the MTVQA Leaderboard (2025), a large-scale multilingual benchmark for multimodal VQA evaluation. We filter models that achieve competitive results on the Vietnamese (VI) subset, as shown in Table 1, and apply them to our custom dataset for consistent cross-evaluation.

4.1. ViBlind-ProductQA dataset

To evaluate fine-grained visual and linguistic comprehension in real-world assistive scenarios, we introduce ViBlind-ProductQA, a human-annotated dataset tailored for Vietnamese product understanding. This dataset reflects practical challenges faced by visually impaired users, such as poor lighting, occlusions, non-standard viewpoints, and diverse packaging designs.

4.1.1. Image Collection and Augmentation

The dataset consists of 75 unique product images, captured under natural conditions with varied backgrounds and angles. To simulate real-world distortions and assess spatial robustness, each image undergoes two augmentations: a horizontal flip combined with a random rotation of +30° and −30°, respectively. This yields 225 images in total (Figures 2 and 3).

Each image is paired with five natural language question–answer (QA) pairs, resulting in 1,125 QA pairs in total. The questions follow the Visual Question Answering (VQA) format and are designed to test both visual recognition (e.g., color, shape) and OCR-based text understanding (e.g., product name, expiration date).

Figure 2. Horizontally flipped and rotated (+30°)

(Note: Image captured by the author)

Figure 3. Horizontally flipped and rotated (-30°)

(Note: Image captured by the authors)

4.1.2. Question-answer annotation

To assess fine-grained visual and textual comprehension, each image in ViBlind- ProductQA is annotated with five natural language question–answer pairs, following the standard Visual Question Answering (VQA) format. These questions are carefully crafted to reflect common and realistic information needs of visually impaired users, such as identifying a product's name, usage, or expiration date. Each QA pair is tagged with a corresponding semantic field, enabling structured evaluation across diverse product-related inquiries.

The five information fields are listed in Table 1, along with example questions in both English and Vietnamese.

This annotation strategy ensures that each image includes a diverse set of question types, supporting both visual recognition (e.g., color, shape) and text

extraction (e.g., product name, expiration date). Across 225 images, the dataset provides 1,125 annotated QA pairs, facilitating detailed and targeted evaluation.

Table 2 provides an example annotation illustrating the structure of a typical QA entry.

Table 1. Information fields and question examples in the dataset

Information Field	Example Question
Product Name - Tên sản phẩm	What is the name of the product in the image? - Tên sản phẩm trong ảnh là gì?
Product Type - Loại sản phẩm	Can you tell me the type of this product? - Có thể cho biết loại của sản phẩm này được không?
Color / Shape - Màu sắc / Hình dáng	Can you describe the color of the product? - Bạn có thể mô tả màu sắc của sản phẩm không?
Usage / Instructions - Công dụng / Hướng dẫn sử dụng	Can you tell me the usage of the product? - Bạn có thể cho biết công dụng của sản phẩm không?

Table 2. Table captions should be placed above the tables, left aligned

Image	Field	Ví dụ câu hỏi
	Image_id	V2_16
	Question_id	1
	Question	What is the name of the product in the image? - Tên sản phẩm trong ảnh là gì?
	Information Field	Product Name - Tên sản phẩm
	True Answer	JOMI
	Model Answer (Gemini)	JOMI Antibacterial Cotton Swabs - Tăm bông kháng khuẩn JOMI

Although ViBlind-ProductQA is relatively small in scale, its focused design, semantic labeling, and real-world relevance make it a valuable benchmark for evaluating multimodal models in Vietnamese product VQA tasks. It enables fine-grained analysis of model outputs across distinct semantic categories and supports practical deployment scenarios, especially for assistive applications like NaviBlind.

4.2. Baseline models and benchmark reference

To build a reliable and effective AI assistant system, selecting a foundational multimodal large language model (VLM) is the most crucial step. Instead of relying on intuition, we conducted a quantitative analysis based on the performance of top models on the reputable Open VLM Leaderboard (OpenCompass, 2024). The goal of this step is to identify the model with the strongest general capabilities before proceeding to detailed evaluation on the Vietnamese dataset.

We considered three top candidates: Qwen-VL-Max, InternVL-Chat-V1.5, and Gemini 2.0 Flash. Their performance across various benchmarks is presented in the table below.

The results from Table 3 clearly show that the Qwen2.5-VL-72B model achieved the highest score (44.8) in the TextVQA task. It not only outperformed its counterpart in the same category, InternVL3-14B, but also demonstrated better performance than the Gemini 1.5 Pro model in this specialized task.

Table 3. Performance comparison of top VLMs on the open VLM leaderboard

Benchmark	Qwen2.5-VL-72B	InternVL3-14B	Gemini 1.5 Pro
TextVQA	44.8	36.2	41.3

Its outstanding ability to read and understand text in images confirms that Qwen2.5-VL-72B is the most suitable candidate for the requirements of the NaviBlind project. Therefore, we decided to select Qwen2.5-VL-72B as the core model for development and integration into the system. The following sections of this paper will present a detailed evaluation and performance analysis of this model on a custom-built Vietnamese dataset designed for the application.

These models were then evaluated on the ViBlind-ProductQA dataset using consistent prompts derived from user speech and standardized settings, allowing for a controlled performance analysis.

4.3. Evaluation metrics

To objectively evaluate the performance of baseline models in our use cases, we designed a quantitative evaluation protocol based on the ViBlind-ProductQA dataset. Each model is provided with a product image and a corresponding question and is expected to generate an accurate textual response in Vietnamese. However, a single question may yield multiple valid responses in Vietnamese, with variations in wording while remaining semantically correct. Therefore, we designed our system to be flexible in expression, ensuring that such variations do not compromise the accuracy of information extraction. This flexibility also helps our chatbot sound more natural instead of only giving one fixed answer based on pre-labeled data.

For this purpose, we evaluate model responses using the following widely accepted metrics for natural language generation tasks:

– ***METEOR (Metric for Evaluation of Translation with Explicit ORdering):*** METEOR goes beyond exact word matching by considering synonyms, stemming, and word order. It calculates a harmonic mean of precision and recall, allowing it to reward semantically similar responses even if the exact words differ (Banerjee & Lavie, 2005)

– ***BERTScore:*** This metric compares the semantic similarity between the model's response and the reference answer by computing the cosine similarity of their contextual embeddings using BERT (Devlin et al., 2019). Unlike traditional metrics, BERTScore captures deeper meaning rather than just surface-level word overlap.

– ***Accuracy:*** While the two metrics above (METEOR and BERTScore) are useful for evaluating answer quality, they can be difficult to interpret in terms of simple correctness. Since BERTScore measures semantic similarity, we adopt a fixed threshold of 0.7: if the similarity score between the model's answer and the human-labeled reference is greater than or equal to 0.7, we consider it a correct answer. Accuracy is then calculated as the ratio of correct answers to the total number of questions. This gives us an intuitive understanding of how often the model produces acceptable responses. For example, if we evaluate the model on 100 questions and 80 of its answers are considered correct under this criterion, the accuracy would be 80%.

Each model is evaluated on all 225 images and 1,125 question-answer pairs in the dataset. For every question, the generated response is compared against a single human-annotated reference answer. To ensure fair and consistent comparison across models, we apply a standardized instruction prompt and fix the temperature parameter to reduce randomness in the output.

By relying exclusively on quantitative metrics, this evaluation offers a rigorous and reproducible assessment of each model's ability to perform Vietnamese product VQA under realistic conditions.

4.4. Results and analysis

Table 4 presents the evaluation metrics of our baseline models on the ViBlind- ProductQA dataset. Overall, all three models perform comparably, with Gemini 2.0 Flash achieving the highest scores across the board.

Table 4. Performance on ViBlind-ProductQA dataset

Metric	Gemini 2.0 Flash	Qwen2.5-VL-72B	InternVL3-14B
METEOR	**0.4572**	0.4044	0.3865
BERT-Score	**0.8497**	0.8344	0.8179
Accuracy	**77.81%**	77.37%	70.06%

Specifically, Gemini reaches the highest METEOR score at 0.4572, followed by Qwen2.5 at 0.4044 and InternVL3 at 0.3865. In terms of semantic similarity, measured by BERTScore, Gemini also leads with a score of 0.85, while Qwen2.5 and InternVL3 achieve 0.83 and 0.82, respectively. These results indicate that Gemini's outputs align most closely with human judgment in both lexical and semantic terms.

Regarding correctness, Gemini answered 894 out of 1,149 questions correctly, achieving an accuracy of 77.81%, which is slightly higher than Qwen2.5's 77.37% and clearly higher than InternVL3's 70.06%.

It is worth noting that this comparison is not strictly about fairness, but rather about evaluating which model best aligns with our task requirements. While Gemini 2.0 Flash outperforms others, it is a closed API, and specific details such as model size remain unclear. In contrast, Qwen2.5 has a publicly known size of 72B parameters, which is significantly larger than InternVL3's 14B, making the comparison more about performance in practice than equal capacity. Given its strong performance in both Table 3 and Table 4, we choose to adopt Gemini 2.0 Flash as the core model in our system, due to its consistent

superiority in both accuracy and semantic evaluation metrics.

5. CONCLUSION

In this paper, we introduced NaviBlind, a multimodal AI assistant that helps visually impaired users access essential product information through voice-driven interaction. By combining PhoWhisper-medium for Vietnamese speech recognition, Gemini 2.0 Flash for image understanding, and gTTS for text-to-speech output, the system allows users to speak their information needs and receive tailored spoken responses without relying on barcodes or tactile cues.

We introduced ViBlind-ProductQA, a custom dataset reflecting real-world product recognition challenges. Quantitative evaluations using METEOR, BERTScore, and Accuracy show that Gemini 2.0 Flash outperforms strong baselines such as Qwen2.5- VL-72B and InternVL3, particularly in generating semantically aligned answers.

Our results highlight the potential of combining vision-language models with speech-based prompts to support assistive technology. NaviBlind demonstrates how human-centric AI can be translated into accessible, impactful technology supporting autonomy and inclusion for visually impaired communities.

REFERENCES

Banerjee, S., & Lavie, A. (2005, June). 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* (pp. 65–72).

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of *the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)* (pp. 4171–4186).

Doan, K. T., Huynh, B. G., Hoang, D. T., Pham, T. D., Pham, N. H., Nguyen, Q. T. M., Vo, B. Q., & Hoang, S. N. (2024). *Vintern-1B: An efficient multimodal large language model for Vietnamese*. arXiv preprint arXiv:2408.12480.

Google. (2018). *Use Lookout to explore your surroundings*. Android Accessibility Help. Google. https://support.google.com/accessibility/android/answer/9031274?hl=en

Google Cloud. (2018). *Speech-to-Text AI: Speech recognition and transcription*. Google. https://cloud.google.com/speech-to-text

Le, T. T., Nguyen, L. T., & Nguyen, D. Q. (2024). *Phowhisper: Automatic speech recognition for vietnamese*. arXiv preprint arXiv:2406.02555.

Li, J., Li, D., Savarese, S., & Hoi, S. (2023, July). Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning* (pp. 19730-19742). PMLR.

Microsoft Garage. (2024). *Seeing AI.* Microsoft. https://www.microsoft.com/en-us/garage/wall-of-fame/seeing-ai/

OpenAI. (2022, September 21). *Introducing Whisper.* https://openai.com/index/whisper/

OpenCompass. (2024). *Open VLM Leaderboard*. https://huggingface.co/spaces/opencompass/open_vlm_leaderboard/

Pndurette. (2025, January 15). gTTS: *Python library and CLI tool to interface with Google Translate's text-to-speech API*. GitHub. https://github.com/pndurette/gTTS

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2023, July). Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning* (pp. 28492-28518). PMLR.

Team, G., Anil, R., Borgeaud, S., Alayrac, J. B., Yu, J., Soricut, R., ... & Blanco, L. (2023). *Gemini: A family of highly capable multimodal models*. arXiv preprint arXiv:2312.11805.

Team, G., Georgiev, P., Lei, V. I., Burnell, R., Bai, L., Gulati, A., ... & Batsaikhan, B. O. (2024). *Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context*. arXiv preprint arXiv:2403.05530.

Tran, C., & Thanh, H. L. (2024). *Lavy: Vietnamese multimodal large language model*. arXiv preprint arXiv:2404.07922.

Be My Eyes. https://en.wikipedia.org/wiki/Be_My_Eyes

Zhu, D., Chen, J., Shen, X., Li, X., & Elhoseiny, M. (2023). MiniGPT-4: Enhancing vision-language understanding with advanced large language models. In *Proceedings of the 12th International Conference on Learning Representations (ICLR 2024).* Open Review.

https://openreview.net/forum?id=1tZbq88f27