Proceedings of the Thirty-Fifth International Joint Conference on Artificial Intelligence (IJCAI-26)
Demonstrations Track

AI-Powered Interactive Multimodal Digital Book & Online Shop for Blind and Visually Impaired Users

Mazen Salous[1], Daniel Westphal[1], Wilko Heuten[1], Ayoub Ben Dhiab[2], Charles Hudin[2], Sabrina Paneels[2], Susanne Boll[3], Larbi Abdenebaoui[1]
[1]OFFIS Institute for Information Technology, Oldenburg, Lower Saxony, Germany
[2]Université Paris-Saclay, CEA, List, F-91120, Palaiseau, France
[3]University of Oldenburg, Oldenburg, Lower Saxony, Germany
mazen.salous@offis.de, Daniel.Westphal@offis.de, wilko.heuten@offis.de, charles.hudin@cea.fr, susanne.boll@informatik.uni-oldenburg.de, larbi.abdenebaoui@offis.de

Abstract

We present an AI-powered interactive multimodal system that enriches digital image accessibility for blind and visually impaired (BVI) users. Our demonstration showcases two application domains: (1) an educational digital book, and (2) an online shopping interface. In both use-cases, users can virtually feel material textures (leather, wood, etc.) and engage in voice-driven inquiry about images. The system integrates state-of-the-art AI components – including voice-to-voice conversational agents, vision models for object segmentation, and a custom 16-actuator vibrotactile display – to provide multimodal feedback (haptic vibrations, spoken descriptions, and audio cues). The result is an inclusive technology with significant societal benefit, empowering BVI users to learn and shop more independently through natural multimodal interactions.

1 Introduction

Digital services in education, entertainment, and shopping increasingly rely on visual online content. Digital images play significant roles in all digital world aspects. However, images and visual media remain largely inaccessible to people who are blind or visually impaired. Traditional approaches for conveying images (e.g. tactile graphics or long alt-text descriptions) are static, labor-intensive, and limited in detail [Wu *et al.*, 2025a]. We investigate two main potentials for improving the digital accessibility: AI-powered interactive and multimodal support. As sighted users can continuously extract additional information by visually inspecting finer details in an image, our system compensates for this advantage through human-centered AI interaction: BVI users explore by touch and voice, while the AI responds with layered multimodal feedback — audio cues, haptic patterns, and spoken descriptions that can evolve into a dialogue.

In this paper, we introduce an interactive system that allows BVI users to touch an image and hear/feel meaningful feedback in real time, while asking an AI visual assistant about the content. By combining computer vision, natural language processing, and haptic technology, our system transforms visual scenes into an audio-tactile experience. We present two use-cases: digital book and online shop. In the digital book scenario, users explore educational images (e.g. in a textbook or story) via a vibrotactile tablet and receive spoken annotations and textured vibrations corresponding to image regions. In the online shopping scenario, users can virtually feel product materials and ask questions (via voice) about item details (colors, patterns, features) as they browse. An OpenAI-based voice assistant enables seamless voice-to-voice interaction, allowing the user to inquire about any visible detail; meanwhile, vision models identify objects and regions on-the-fly to provide relevant answers. The system emphasizes AI for society by improving information access for BVI individuals, and exemplifies AI interaction through its integration of multiple AI modalities into a cohesive interactive experience.

Our work builds on prior research in multimodal accessibility. For example, audio-tactile presentation of graphics has been shown to improve comprehension for BVI learners [Sun *et al.*, 2025]. Refreshable tactile displays paired with audio narration (e.g. "tactile data comics") significantly enhance understanding of visual data by blind students [Sun *et al.*, 2025]. Likewise, multimodal educational tools like TaleVision use AI-generated tactile graphics and sounds to aid blind children's learning [Wu *et al.*, 2025a]. However, existing solutions often lack interactivity or real-time adaptability. Our system pushes the boundary by enabling real-time, interactive exploration: users can physically navigate images and converse with an AI about any sub-part of the scene. This dynamic "dialogue with the image" is made possible through cutting-edge segmentation models and large language models (LLMs). In the following, we detail the problem scenario motivating our system, the application domains, the underlying technology, and the innovative aspects of our demo. We also describe the interactive experience we will provide, and discuss the broader impact on society.

2 Related Works

Recent work has increasingly focused on multimodal approaches to make visual content accessible to blind and visually impaired (BVI) users. Audio-tactile tools such as Tactile Data Comics [Sun *et al.*, 2025] and TaleVision [Wu *et al.*,

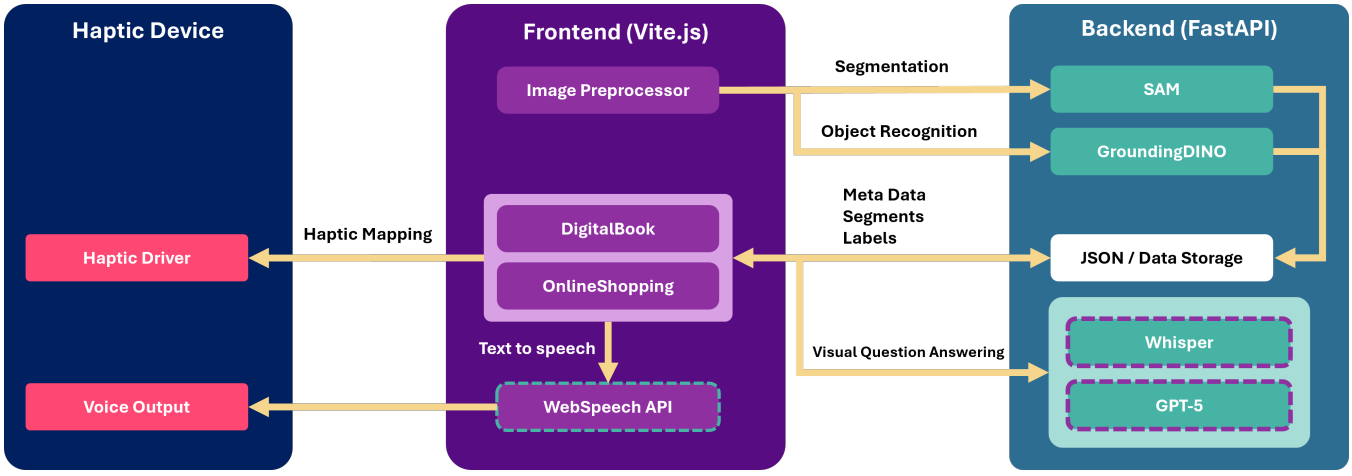

Figure 1: System architecture of our multimodal interface, showing the integrated components between frontend, backend, AI services, and haptic hardware.

2025a] show how synchronizing tactile displays with narrated feedback enhances educational comprehension and supports interactive learning. These systems demonstrate the value of combining touch and audio, yet often rely on scripted content and lack live, user-driven exploration. Our system complements these efforts by introducing a conversational agent that dynamically answers user queries during tactile interaction, transforming static exploration into an interactive dialogue. Similarly, vibrotactile hardware like Vipins [Wu *et al.*, 2025b] and multimodal tablets [Maćkowski *et al.*, 2023] have demonstrated how vibratory cues can support shape and texture recognition. We extend this line with a cost-efficient 16-actuator tactile tablet that maps image semantics to vibrotactile patterns, drawing on prior insights into texture simulation using electrotactile stimuli [Germani *et al.*, 2013] and our own haptic rendering work [Salous *et al.*, 2026].

Beyond passive output, our system integrates conversational AI and computer vision into real-time touch-based interaction. This moves past static image captioning—typical of systems like Seeing AI—toward an open-ended, voice-first interface where users inquire about sub-regions and materials within images. While prior studies show vibrotactile cues can evoke material qualities [Germani *et al.*, 2013], [Wu *et al.*, 2025b], we link these to live image segmentation and dialogue. Our system unites segmentation models with large language models to enable sub-object identification and voice-driven queries (”Is this sleeve leather or cotton? - Touch and Feel it!”), complementing past work with a uniquely adaptive and context-aware user experience for BVI shopping and education.

3 Problem Scenarios

Digital Book: Our system addresses the challenge of inaccessible imagery in education for BVI learners. Traditional tactile graphics or verbal descriptions are static and limited, often requiring external assistance. By contrast, our solution enables students to explore educational visuals—such as diagrams or illustrations—independently through a vibrotactile tablet that renders textured boundaries and provides spoken labels. Crucially, users can ask follow-up questions via voice (e.g., “Tell me more about Saturn”), allowing for dynamic, contextual learning through conversational AI. This supports equal access to complex visual content and fosters curiosity-driven, self-paced education.

Online shopping: BVI users face barriers in understanding product materials and visual details. Our system transforms product imagery into touchable textures—e.g., simulating denim or leather—allowing users to distinguish regions and query specific attributes by voice. A shopper might touch a sleeve and ask, “Is this denim?” or “Is there a softer version?” receiving intelligent responses based on segmentation and product metadata. These scenarios demonstrate the core motivation: to deliver real-time, multisensory access to visual content by fusing touch, speech, and vision AI—ultimately enhancing independence, confidence, and inclusion across everyday digital experiences.

4 System Overview

Our system as shown in Figure 1 integrates four components—vision, language, haptics, and interaction—into a cohesive multimodal interface. The vision module combines SAM [Kirillov *et al.*, 2023] for touch-triggered segmentation and GroundingDINO [Liu *et al.*, 2023] for open-set object recognition, enabling dynamic identification of sub-objects as users explore. A voice-to-voice dialogue module powered by Whisper and GPT-4 translates user queries into contextual responses, referencing touch position and visual data. This creates a fluid conversational loop where users can touch any region and ask spontaneous questions about it, with the system tracking and adapting to the interaction flow.

Haptic rendering is delivered via a 15.6-inch OLED touch display with 16 rear-mounted piezoelectric actuators, providing localized vibrotactile feedback. Haptic patterns from our tactile pattern library combine temporal waveforms and spatial masks: segmented image regions trigger material-specific waveforms characterized by frequency, amplitude, modula-

tion, and envelope [Salous *et al.*, 2026], enabling dynamic mapping to object textures and boundaries [Wu *et al.*, 2025b]. The design emphasizes intuitive "touch-and-talk" use [Edirisinghe *et al.*, 2022], offering real-time, low-latency feedback in a portable and accessible form factor for educational and consumer contexts.

5 Interactive Experience

During the live demo, attendees will explore two real-time scenarios using our vibrotactile tablet and voice assistant. In the digital book setting, users can touch and query parts of an animal illustration—such as feathers or wings—and receive multimodal feedback: a vibration pattern signals the region, and the AI assistant verbally identifies and describes it. As shown in Figure 2, touching the bird's wing triggers both haptic feedback and spoken labels, and users may ask follow-ups like "What am I touching now?", receiving contextual answers instantly.

Figure 2: Touch-based exploration of segmented sub-parts in a bird illustration.

In the online shop demo, users explore materials on a product image—such as a sneaker—and feel distinct textures (e.g., fabric vs. logo patch), as in Figure 3. Touching the logo yields a sharp, smooth vibration, while other regions simulate coarser or softer materials. Voice queries like "Is this rubber or fabric?" prompt intelligent, localized responses. These scenarios demonstrate how blind users can interactively access visual and material information through seamless AI-driven multimodal feedback.

6 Innovations

Our system introduces a unique integration of vision, language, and haptics for image accessibility, enabling blind users to explore digital content interactively through touch and dialogue. Unlike prior solutions limited to static descriptions or tactile graphics, we provide a synchronized experience where users can feel textures, hear semantic labels, and engage in contextual conversation with an AI assistant. This real-time, multimodal feedback loop bridges physical interaction with AI reasoning—an embodiment of *AI Systems & Interaction* that empowers nonvisual image understanding.

Technically, we advance assistive interaction through on-the-fly segmentation and open-set object discovery using SAM [Kirillov *et al.*, 2023] and GroundingDINO [Liu *et al.*, 2023], allowing users to explore not only known objects but also their sub-parts without pre-annotation. Conversational AI extends beyond captioning to support clarification, follow-ups, and image-specific dialogue. Our haptic module complements this by rendering distinct vibrotactile patterns for visual regions and materials—applying prior findings on tactile icons [Germani *et al.*, 2013], [Wu *et al.*, 2025b] in a deployable, low-cost device tested with AI-powered haptic generation pipelines [Salous *et al.*, 2026].

Figure 3: Exploring material properties in an online shop interface.

Finally, the system adapts to user preferences in real time: minimizing speech output for tactile-first users or offering richer verbal guidance when needed. This personalization supports inclusive design for varied user strategies and levels of experience. By uniting accessible hardware, AI-driven perception, and conversational interaction, our demo showcases a scalable solution with clear societal impact.

7 Conclusion

We introduced an AI-powered multimodal system that enables blind and visually impaired users to explore visual content through touch, audio, and conversation. By combining real-time segmentation, voice-based interaction, and vibrotactile feedback, our demo addresses key accessibility barriers in education and online shopping, illustrating the potential of *AI Systems & Interaction* to support *AI for Society*.

Built from open-source and low-cost components, the system is practical for future deployment in schools, libraries, or homes. Early user feedback indicates the promise of the approach: users could explore at their own pace, ask contextual questions, and navigate visual information more independently. We provide our demo video under the following link:

https://www.veed.io/view/
5c4bf02e-e864-48bd-bffa-93b1fb987c44?panel=share,

Acknowledgments

This work is part of the European project ABILITY, funded by European Commission under Grant number: 101070396.

References

[Edirisinghe *et al.*, 2022] Chamari Edirisinghe, Norihidayati Podari, and Adrian David Cheok. A multi-sensory interactive reading experience for visually impaired children: a user evaluation. In *Personal and Ubiquitous Computing*, volume 26, pages 807–819, 2022.

[Germani *et al.*, 2013] Michela Germani, Donatella Branciari, Luca Cornelissi, and Mauro Beccaluva. Electrotactile device for material texture simulation. *International Journal of Advanced Manufacturing Technology*, 67:1853–1861, 2013.

[Kirillov *et al.*, 2023] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023.

[Liu *et al.*, 2023] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023.

[Maćkowski *et al.*, 2023] Michał Maćkowski, Piotr Brzoza, Mateusz Kawulok, Rafał Meisel, and Dominik Spinczyk. Multimodal presentation of interactive audio-tactile graphics supporting the perception of visual information by blind people. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 19(2):67:1–67:19, 2023.

[Salous *et al.*, 2026] Mazen Salous, Matthias Karmer, Wilko Heuten, Charles Hudin, Susanne Boll, and Larbi Abdenebaoui. Augmenting imagery with multimodal vibrotactile representations: Touch, feel, and hear. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*, CHI '26, New York, NY, USA, 2026. Association for Computing Machinery. Author preprint / accepted version.

[Sun *et al.*, 2025] Ruoting Sun, Rong Luo, Xiwen Yao, Xinran She, Kotaro Hara, and Yang Jiao. Tactile data comics: Combining step-by-step presentation of tactile graphics with verbal narration for the blind and visually impaired. In *Extended Abstracts of the 2025 CHI Conference on Human Factors in Computing Systems (CHI EA '25)*, pages 1–8, 2025.

[Wu *et al.*, 2025a] Hantian Wu, Hongyi Yang, Fangyuan Chang, Dian Zhu, and Zhao Liu. Ai-generated tactile graphics for visually impaired children: A usability study of a multimodal educational product (talevision). *International Journal of Human-Computer Studies*, 200:103525, 2025.

[Wu *et al.*, 2025b] Qian Wu, Jianguang Li, Hasti Seifi, and Kasper Hornbæk. Vipins: Combining a pin array and vibrotactile actuators to render complex shapes and textures. *International Journal of Human-Computer Studies*, 197:103464, 2025.