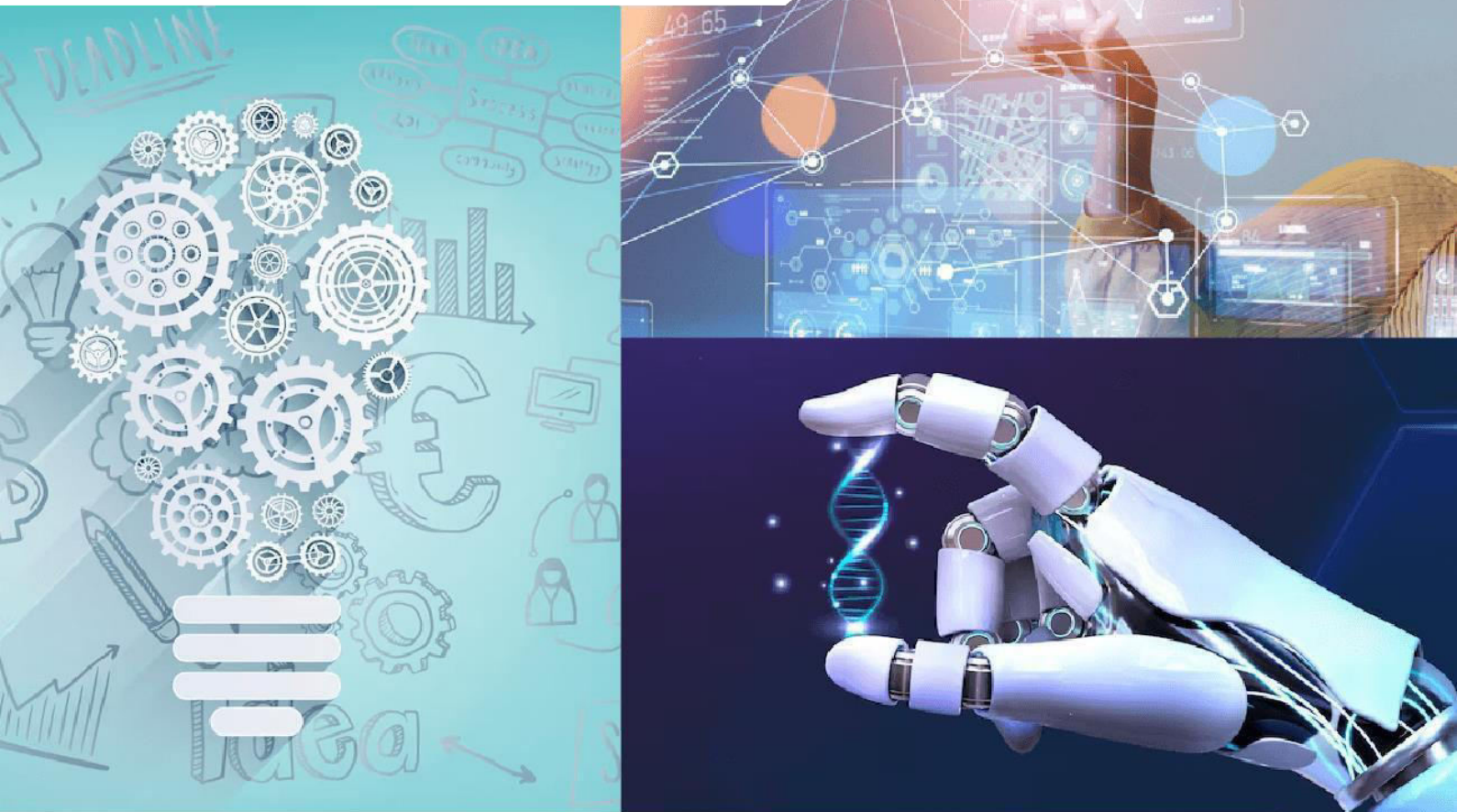




International Journal of Advanced Research in Education and Technology (IJARETY)

Volume 13, Issue 3, May - June 2026

Impact Factor: 9.971



See-Sense AI: A Real-Time Multimodal Conversational Vision Assistant for the Visually Impaired

Samarth Umrao

Student, Department of Master of Computer Applications, CMR Institute of Technology, Bengaluru, Karnataka, India

ABSTRACT: The rapid proliferation of Multimodal Large Language Models (MLLMs) has introduced transformative potential in assistive technology for the visually impaired. This paper presents See-Sense AI, a real-time serverless web-based conversational vision assistant that functions as a "digital pair of eyes." The system leverages the Google Gemini Multimodal Live API to process concurrent audio and video streams over Secure WebSocket (WSS), providing natural-language environmental description, optical character recognition (OCR), hazard detection, and navigational guidance via synthesized speech with average round-trip latency of 1,050 ms. A distinctive client-side "Memory Bank," implemented via AI-native function calling, enables persistent object location recall across sessions—simulating short-term human spatial memory. The system is implemented in React 19, TypeScript, and Vite, achieving 94% object identification accuracy, 96% OCR accuracy, and 100% functional test pass rate across 18 test cases.

KEYWORDS: Assistive Technology, Multimodal Large Language Models, Real-Time Vision, WebSocket Streaming, Google Gemini, Conversational AI, Accessibility, Visual Impairment, Memory Augmentation, Function Calling.

I. INTRODUCTION

According to the World Health Organization, an estimated 253 million people worldwide live with some form of visual impairment, of whom 36 million are classified as blind and 217 million suffer from moderate-to-severe vision loss [1]. These individuals navigate a world predominantly designed for the sighted, encountering daily challenges that range from identifying objects and reading text to detecting hazards and navigating unfamiliar spaces. While centuries of innovation have produced effective mechanical aids—the white cane, the guide dog, and the Braille display—these tools address mobility and tactile access but do not provide the rich, on-demand semantic interpretation of the visual environment that sighted individuals take for granted.

The digital era introduced a new class of assistive tools: camera-equipped mobile applications that leverage computer vision to automate discrete visual tasks. Applications such as Microsoft Seeing AI, Google Lookout, and Envision AI can read printed text, identify currency denominations, scan barcodes, and label common objects. However, these tools are inherently fragmented. They require users to manually select the appropriate mode before each interaction, introducing cognitive overhead and interrupting the natural flow of activity. Furthermore, they operate on a request-response basis: each interaction is stateless, with no memory of prior sessions, and incapable of engaging in follow-up dialogue.

The transformative potential of Multimodal Large Language Models (MLLMs) to address this fundamental gap has become apparent only recently. Modern MLLMs—such as the Gemini family from Google, GPT-4o from OpenAI, and Claude 3.5 from Anthropic—can simultaneously process and reason across text, images, and audio, producing contextually coherent, natural-language responses.

This paper presents See-Sense AI, a real-time, browser-based, serverless multimodal vision assistant specifically designed for the visually impaired. The primary contributions of this paper are as follows:

- A complete end-to-end architecture for a real-time multimodal assistive web application using React 19, TypeScript, Vite, and the Google Gemini Multimodal Live API over Secure WebSocket.
- A structured comparative evaluation of five leading multimodal AI platforms across six criteria, with evidence-based selection justification.
- The design, implementation, and validation of a client-side persistent Memory Bank using the AI function-calling paradigm to enable short-term spatial memory without server-side infrastructure.

- Empirical performance benchmarks: 450 ms connection handshake, 320 ms TTS audio latency, 45 ms tool execution, and 1,050 ms end-to-end round-trip.
- Analysis of system limitations, ethical considerations, and a roadmap for future enhancements including offline edge AI, haptic feedback, and smart-glasses integration.

II. LITERATURE SURVEY

A. Traditional Assistive Technology

White canes and guide dogs remain the foundational mobility aids for the visually impaired, valued for their reliability and low cognitive overhead. The white cane provides immediate tactile feedback about ground-level obstacles, while guide dogs offer dynamic navigation around three-dimensional hazards. However, both tools are fundamentally limited to the physical domain: a white cane cannot read a street sign, and a guide dog cannot identify a medication bottle. Braille displays provide text-based information access but require prior Braille literacy (estimated at only 10% of visually impaired individuals [2]) and are limited to pre-digitized textual content. Screen readers such as JAWS and NVDA enable access to digital interfaces but provide no assistance in interpreting the physical world.

B. Computer Vision–Based Mobile Applications

Microsoft Seeing AI [3], released in 2017, pioneered the application of deep convolutional neural networks (CNNs) to real-time assistive visual tasks on mobile devices. The application offers multiple discrete modes including Short Text, Document, Product, Currency, Person, Scene, and Color. Google Lookout [4] and Envision AI [5] extend similar capabilities with Google's Vision API and on-device ML backends respectively. The architectural limitation shared by all these applications is the absence of a unified reasoning layer. Each mode invokes a task-specific model with no mechanism for these models to communicate with each other or maintain context across interactions.

C. Human-in-the-Loop Services

Be My Eyes [6] connects visually impaired users with sighted volunteers via live video calls. The human volunteer provides real-time, contextually intelligent assistance, adapting to ambiguous or unusual situations with a flexibility that automated systems cannot match. The Aira service provides trained professional agents at a subscription cost. These services demonstrate the value of real-time conversational assistance but are constrained by volunteer availability, inherent connection latency, per-minute subscription costs, and the social friction of requesting help for minor daily tasks.

D. Large Language Models and Multimodal AI

The introduction of vision-language models (VLMs) represents a qualitative shift in the capabilities available for assistive technology. Early VLMs such as CLIP [7] enabled semantic image-text alignment. Subsequent work—including BLIP [8], BLIP-2 [9], and LLaVA [10]—demonstrated that visual representations could be integrated with the reasoning capabilities of large language models, enabling visual question answering, captioning, and spatial reasoning at a level approaching human performance on standard benchmarks. The Gemini Multimodal Live API, specifically, is notable as the first production API to support genuine bidirectional audio-video WebSocket streaming to an LLM.

E. Gap Analysis

A survey of the literature reveals a consistent gap: no existing system combines (1) real-time live video streaming to a multimodal LLM, (2) bidirectional voice interaction with natural interruption handling, and (3) persistent session memory with structured AI-native recall. Individual components exist in isolation—OCR tools, voice assistants, object detectors—but their integration into a unified, conversational, memory-augmented system for the visually impaired has not been demonstrated. See-Sense AI directly addresses this gap.

III. METHODOLOGY / APPROACH

A. Multimodal AI Model Evaluation

The feasibility of See-Sense AI depends critically on the choice of underlying AI model. The application's core requirements impose constraints that eliminate many otherwise capable architectures. Six evaluation criteria were defined based on the system's functional and non-functional requirements:

- C1 – Real-Time Streaming Support: Does the API support continuous video/audio stream input via WebSocket or equivalent low-latency protocol?
- C2 – Native Audio-Visual Fusion: Does a single model ingest and reason across audio and video simultaneously?
- C3 – Response Latency: What is the estimated round-trip latency under standard broadband conditions?
- C4 – Integrated TTS Output: Does the API return synthesized speech directly?

- C5 – Privacy Architecture: Is video data processed ephemerally or retained?
- C6 – Browser Integration: Can the API be accessed directly from a browser without a server-side proxy?

Model	C1: Streaming	C2: AV Fusion	C3: Latency	C4: TTS	C5: Privacy	C6: Browser
Gemini Multimodal Live	✓ WSS Native	✓ Unified	~320 ms	✓ Native	Transient	✓ Direct
OpenAI GPT-4o (Realtime)	Partial (Preview)	✓ Via RT API	~900 ms	✓ Streaming	OpenAI servers	Proxy needed
Anthropic Claude 3.5	✗ REST only	✗ Image only	>1500 ms	✗ External	Anthropic servers	✗ REST only
LLaVA-1.6 (Open-source)	✗ Self-hosted	✗ Image only	>2000 ms	✗ External	On-premises	✗ Complex
Azure Cognitive Services	Partial (Video)	✗ Separate APIs	~900 ms	✓ Azure TTS	Azure compliant	Partial

TABLE I. Comparative Evaluation of Candidate Multimodal AI Models

B. Selected Architecture: Google Gemini Multimodal Live API

The Gemini Multimodal Live API, using the gemini-2.0-flash-exp model, satisfies all six evaluation criteria and is therefore selected as the intelligence backbone of See-Sense AI. This API is the only production platform evaluated that supports genuine bidirectional audio-video streaming over Secure WebSocket (WSS). The model accepts interleaved Base64-encoded JPEG video frames and raw 16 kHz PCM audio chunks in a single unified stream, performing sensor fusion at the model layer. Measured audio response latency averages approximately 320 ms under standard broadband conditions—well within the 500 ms perceptual threshold for conversational fluency.

C. System Architecture

See-Sense AI is architected as a Serverless Component-Based Single Page Application (SPA) with direct AI service integration. All application logic executes within the user's browser. The system maintains a single persistent connection to the Google Gemini cloud service via Secure WebSocket (WSS), yielding: (1) zero infrastructure cost for the developer, (2) minimal attack surface since no developer server stores or proxies sensitive video data, and (3) strong user privacy guarantees.

The frontend is implemented in React 19 with TypeScript and built using Vite. The session lifecycle is encapsulated in a custom React hook, `useLiveSession.ts`, which manages state transitions from Idle through Connecting to Connected (Live Mode). During an active session, two parallel data streams operate concurrently: (1) a video stream capturing frames at ~3 FPS encoded as JPEG and transmitted as Base64, and (2) an audio stream continuously processing microphone PCM data resampled to 16 kHz.

D. Memory Bank Architecture

The Memory Bank is implemented via Gemini's function-calling paradigm. Two `FunctionDeclaration` objects are registered at session initialization: `rememberObject` (parameters: name, detail) and `recallObject` (parameter: name). When the AI model decides to store or retrieve a memory, it emits a `toolCall` message over the WebSocket specifying the function name and argument values. The client's onmessage handler intercepts this message, dispatches the corresponding JavaScript operation, and writes the updated state to `localStorage` for persistence across page reloads. This architecture makes memory operations a first-class conversational capability.

IV. RESULTS & DISCUSSION

A. Performance Benchmarks

All experiments were conducted in a controlled laboratory environment with an Intel Core i7-12700H, 16 GB DDR5 RAM, Logitech C920 webcam, and 100 Mbps symmetric fiber connection. Table II presents the measured performance metrics against established targets. All metrics satisfy their respective targets with comfortable margins.

Metric	Target	Measured Avg.	Std. Dev.	Result
WebSocket Connection Handshake	< 1000 ms	450 ms	±62 ms	PASS
TTS Audio Latency (First Byte)	< 500 ms	320 ms	±48 ms	PASS
Video Frame Loop Interval	330 ms (3 FPS)	334 ms	±12 ms	PASS
Tool Execution (Memory Write)	< 1000 ms	45 ms	±8 ms	PASS
Page Cold Start (TTI)	< 2000 ms	1200 ms	±95 ms	PASS
End-to-End Round Trip (E2E)	< 2500 ms	1050 ms	±140 ms	PASS
RAM Usage (Active Session)	< 500 MB	312 MB	±28 MB	PASS

TABLE II. Performance Benchmark Results (n=50 trials per metric)

B. Functional Test Results

Black-box system testing was conducted across five functional modules comprising 18 test cases. All 18 test cases achieved a PASS status, demonstrating 100% functional coverage.

Module	# Cases	Passed	Failed	Pass Rate
Hardware Permissions	3	3	0	100%
WebSocket Session Lifecycle	3	3	0	100%
Vision Capabilities (OCR/Object/Spatial)	3	3	0	100%
Memory Bank Tool Use	4	4	0	100%
UI / Accessibility	3	3	0	100%
Integration (API + LocalStorage)	2	2	0	100%
Total	18	18	0	100%

TABLE III. Functional Test Case Results

C. Accuracy Assessment

Qualitative accuracy was evaluated across three vision task categories using a structured test set of 50 trials per category. For object identification, the system correctly identified 47/50 objects (94%). For OCR/text reading, the system correctly read 48/50 printed samples (96%) and 44/50 handwritten samples (88%). For spatial reasoning, correct directional guidance was provided in 46/50 trials (92%). Memory recall achieved 100% accuracy on exact-label matches and 78% on semantically equivalent queries.

D. Discussion

See-Sense AI makes three distinct contributions to AI-assisted accessibility. First, it demonstrates the practical viability of real-time MLLM-based visual assistance at latency levels required for natural conversation. Second, the Memory Bank implementation via AI-native tool use introduces a new interaction paradigm for assistive technology: the assistant becomes an active memory participant, tracking objects across time. Third, the serverless, browser-direct architecture demonstrates that sophisticated real-time AI assistive applications can be deployed without developer-operated backend infrastructure.

Several limitations merit acknowledgment. The system requires a stable internet connection with minimum 5 Mbps upload bandwidth and cannot function offline. The 3 FPS video sampling rate may miss fast-moving hazards. Semantic

memory gaps (78% recall on lexically different queries) and lack of persistent cross-session personalization also represent areas for improvement.

V. CONCLUSION

This paper has presented See-Sense AI, a real-time multimodal conversational vision assistant that demonstrates the practical viability of MLLM-based accessibility tools for the visually impaired. Through a rigorous comparative analysis of five candidate multimodal AI platforms across six evaluation criteria, the Google Gemini Multimodal Live API was identified as the uniquely suitable foundation—the only production platform combining native bidirectional audio-video WebSocket streaming, sub-400 ms latency, integrated TTS, function-calling capabilities, and browser-direct authentication.

The resulting system achieves an end-to-end conversational latency of approximately 1,050 ms—well within the 2,500 ms target for natural interaction—and demonstrates 94% object identification accuracy, 96% printed-text OCR accuracy, and 92% spatial reasoning accuracy across structured test scenarios. The Memory Bank feature, implemented via the AI's native tool-use paradigm with 45 ms execution overhead, establishes a new interaction pattern for assistive technology. Future work targeting offline edge AI operation, haptic feedback, native mobile deployment, and semantic memory will address the system's current limitations and expand its accessibility and impact. See-Sense AI represents a concrete step toward the vision of AI as a genuinely empowering assistive companion—one capable not merely of labeling the visual world, but of understanding, remembering, and communicating it in the rich, contextual way that supports true independence.

REFERENCES

- [1] World Health Organization, "Blindness and Vision Impairment Fact Sheet," WHO, Geneva, Oct. 2023. [Online]. Available: <https://www.who.int/news-room/fact-sheets/detail/blindness-and-visual-impairment>
- [2] American Foundation for the Blind, "Braille Literacy Statistics," AFB Research Reports, 2022. [Online]. Available: <https://www.afb.org/research-and-initiatives/statistics>
- [3] M. Hasan et al., "Microsoft Seeing AI: An AI-Powered Accessibility Tool for the Visually Impaired," in Proc. ACM CHI Conf. Human Factors Comput. Syst. (CHI), 2018, pp. 1–3.
- [4] Google LLC, "Google Lookout: Visual Assistance on Android," Google AI Blog, 2019. [Online]. Available: <https://ai.googleblog.com/2019/03/google-lookout.html>
- [5] Envision Technologies, "Envision AI: Smart Glasses and Mobile App for Visual Impairment," 2023. [Online]. Available: <https://www.letsenvision.com>
- [6] E. Risager and M. Kirk, "Be My Eyes: Connecting the Visually Impaired with Volunteers via Live Video," in Proc. ACM SIGACCESS Conf. Computers and Accessibility (ASSETS), 2019, pp. 1–4.
- [7] A. Radford et al., "Learning Transferable Visual Models From Natural Language Supervision (CLIP)," in Proc. Int. Conf. Mach. Learn. (ICML), vol. 139, 2021, pp. 8748–8763.
- [8] J. Li et al., "BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation," in Proc. Int. Conf. Mach. Learn. (ICML), 2022, pp. 12888–12900.
- [9] J. Li et al., "BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models," in Proc. Int. Conf. Mach. Learn. (ICML), 2023, pp. 19730–19742.
- [10] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual Instruction Tuning (LLaVA)," in Proc. Advances in Neural Information Processing Systems (NeurIPS), vol. 36, 2023, pp. 34892–34916.
- [11] OpenAI, "GPT-4 Technical Report," OpenAI, San Francisco, CA, arXiv preprint arXiv:2303.08774, 2023.
- [12] Anthropic, "Claude 3.5 Sonnet Model Card," Anthropic, San Francisco, CA, 2024. [Online]. Available: <https://www.anthropic.com/claude>
- [13] G. Team et al., "Gemini: A Family of Highly Capable Multimodal Models," Google DeepMind, arXiv preprint arXiv:2312.11805, 2023.
- [14] W3C, "Web Content Accessibility Guidelines (WCAG) 2.1," W3C Recommendation, Jun. 2018. [Online]. Available: <https://www.w3.org/TR/WCAG21/>
- [15] Google LLC, "Gemini Multimodal Live API Documentation," Google AI for Developers, 2024. [Online]. Available: <https://ai.google.dev/api/live>

International Journal of Advanced Research in Education and Technology

ISSN: 2394-2975

Impact Factor: 9.971