AI in the Loop: A Multimodal Assistant for Real-Time Lecture Comprehension

Amanda Lacy, Maria Teleki, Esau Hutcherson, Jobin Varughese, Jun Kwon
Frank Shipman, Tracy Hammond
{aklacy,mariateleki,esau.hutcherson,jobinv,jhnkwon,shipman,hammond}@tamu.edu
Texas A&M University
College Station, Texas, USA

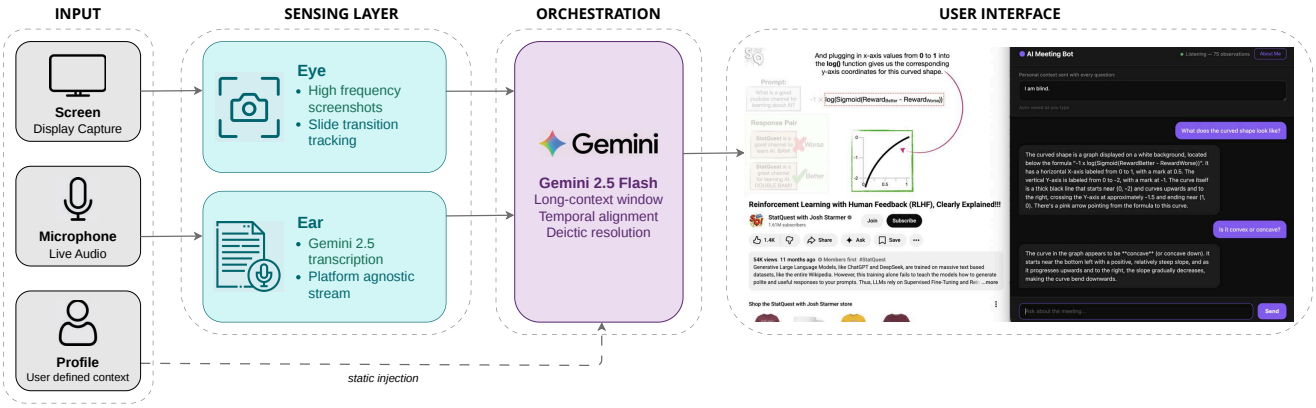

Figure 1: System architecture. The *Eye* and *Ear* sensing modules stream live multimodal input to the long-context Bot orchestrator (`gemini-2.5-flash` [6, 8]), which maintains a persistent session exposed through a minimal browser-based UI. A user profile injects static accessibility context with each query. Demonstrated here with a machine learning lecture video [18].

Abstract

Synchronous learning environments often present a bandwidth problem where high-density visual and auditory information is lost due to fast-paced delivery or inaccessible presentation methods. We present a multimodal assistant designed to provide real-time scaffolding for participant comprehension. The system captures local screen content and audio to maintain a persistent AI session, allowing users to query a long-context model about live content without disrupting the conversational floor. This demo at Learning at Scale invites attendees to interact with the assistant during technical presentations, showcasing its ability to resolve visual ambiguities and summarize recent discussions for improved accessibility, cognitive support, and personalized adaptation to individual access needs.

CCS Concepts

• **Applied computing** → **Interactive learning environments**; • **Human-centered computing** → **Accessibility systems and tools**; *Natural language interfaces*; *Accessibility technologies*; • **Computing methodologies** → Computer vision; Speech recognition.

Keywords

CSCL, Educational Technology, Artificial Intelligence, Digital Accessibility, Synchronous Learning

ACM Reference Format:
Amanda Lacy, Maria Teleki, Esau Hutcherson, Jobin Varughese, Jun Kwon and Frank Shipman, Tracy Hammond. 2026. AI in the Loop: A Multimodal Assistant for Real-Time Lecture Comprehension. In *Learning @ Scale 2026 (L@S '26), June 30–July 03, 2026, Seoul, South Korea.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 Introduction

Modern collaborative software frequently leaves participants in a state of information decay if they miss a fleeting visual cue or a fast-paced comment, a phenomenon recognized in educational psychology as the transient information effect [22]. This "interruption tax" [1] is particularly high for blind or neurodivergent participants [2, 7], non-native speakers [16], or those experiencing high cognitive load – populations for whom synchronous learning environments often fail to provide multiple means of representation [4]. Moreover, these barriers are not uniform: a blind participant and a non-native speaker face qualitatively different information gaps from the same presentation, suggesting that effective support must be personalized to the individual [14, 20, 23, 24].

L@S '26, Seoul, South Korea

ACM ISBN 978-x-xxxx-xxxx-x/YYYY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

To address this, we developed a platform-agnostic assistant that acts as a private, non-visual channel for information retrieval. By offloading clarification queries to an automated system, participants can gain granular control over the meeting flow—reviewing transcripts line-by-line or querying visual data—without signaling confusion to colleagues or consuming shared time. We make the code available at https://anonymous.4open.science/r/ai-bot-8A52.

2 Related Works

Recent advancements in speech recognition and large unified multimodal models have enabled the creation of robust video question-answering (QA) systems. Foundational architectures such as Flamingo utilize gated cross attention layers to improve text-vision embedding alignment [3]. Frameworks such as LLaVA-NEXT have enhanced fine-grained image understanding by using the AnyRes technique to handle processing high resolution visual data [11]. These models have enabled more efficient translation of visual information into text which supports high quality zero-shot question answering.

Previous research indicates that including human-in-the-loop interactions in LLM generated meeting summaries increased user engagement and trust in the system's output [21]. Building on this, we enable users to define the scope of their question, allowing for real-time personalized information generation.

Empirical research suggests that highly automated LLM summarization generation can reduce cognitive engagement. In contrast, intermediate AI solutions improve information retention while reducing cognitive load [5]. Our system is designed to stimulate active cognition by encouraging users to query the AI assistant for additional assistant at the point of information need, rather than generating lengthy post hoc summaries. Numerous social factors, such as aversion to disruption, have been shown to increase cognitive load for neuro-divergent learners in online learning environments [10]. Our system acts as a private assistant to help those who have questions but fear disrupting an on going teaching session to elucidate them. We utilize these previous works as a foundation for our research; however, we differ by allowxing for real-time question answering and personalized information retrieval from a learning environment.

3 System Architecture

The assistant is organized around three modular "senses" that feed into a central intelligence, grounded in the principles of *cognitive load theory* [13, 19] to minimize extraneous processing:

(i) The Eye serves as the visual input channel, capturing high-frequency screenshots of the user's display to track slide transitions and shared documents.

(ii) The Ear handles auditory input by processing raw audio natively via the `gemini-2.5-flash` model [6, 8], enabling the system to categorize environmental sounds and non-verbal cues alongside linguistic transcription while maintaining a platform-agnostic stream through local hardware discovery. This dual-modality approach provides redundant channels of information [9, 12] consistent with Universal Design for Learning principles [4] and shown to support sensory and cognitive access needs.

(iii) The Profile maintains a persistent user profile encoding accessibility preferences and interaction history, enabling the system to adapt its output to the individual — for example, generating rich visual descriptions for a blind user, as shown in Figure 1. This reflects a growing recognition that personalization is itself an accessibility strategy [20, 24].

(iv) The Bot acts as the central orchestrator, utilizing the long-context window of `gemini-2.5-flash` [6, 8] to maintain a continuously updated representation of the session. This architecture allows the system to perform temporal reasoning across modalities, such as resolving deictic references like "that chart mentioned a moment ago" by aligning transcript timestamps with the corresponding visual state.

The system utilizes a standalone browser-based design to bypass platform-specific SDK limitations, ensuring it functions across Zoom, Teams, or local presentations. The interface is deliberately minimal to support cognitive accessibility, featuring only a live transcript panel and a natural-language Q&A field. This "quiet" design philosophy ensures the tool does not add to the visual clutter of modern meeting applications, providing responses only when explicitly prompted by the user.

4 Demonstration Plan

During the Learning at Scale demonstration, attendees will interact with the bot in a simulated technical presentation environment.

4.1 A "Pointing-Heavy" Technical Video

In one scenario, a participant may encounter a "pointing-heavy" technical video, such as a circuit-building tutorial where the instructor references specific pins without verbalizing their coordinates. Such demonstrative pronouns (e.g., "this") are a primary source of confusion for users without visual access [17]. The participant can type a query such as *"Which component is the speaker pointing to right now?"* The Bot retrieves the most recent screenshot, identifies the instructor's hand position relative to the breadboard components, and provides a text-based description [15] that grounds the spoken instruction in the visual context.

4.2 A Mid-Demo Interruption

A second scenario demonstrates cumulative catch-up for participants who approach the demo mid-presentation. Instead of waiting for a natural break in the talk, the attendee can ask the bot to *"Summarize the key points discussed in the last three minutes."* The system interprets the conversational context across multiple transcript segments to produce a coherent summary. This allows the attendee to synchronize with the live discussion immediately, demonstrating the tool's utility as a private resource that reframes the act of seeking clarification from a social cost to an on-demand resource.

5 Future Work

Future work will focus on improving the fidelity of the visual-auditory alignment, expanding the expressiveness of user profiles, and integrating visual emotion recognition to provide participants with nuanced insights into the affective climate of the lecture, such as audience engagement or speaker sentiment. Additionally, we intend to explore how honest expressions of system uncertainty can

further reduce the risk of misinformation in high-stakes educational settings.

6 Conclusion

Our demo illustrates how a multimodal AI assistant can bridge the informational gaps inherent in synchronous collaborative learning. Critically, the system treats personalization as an accessibility strategy in its own right, adapting its output modality and detail level to individual profiles.

References

[1] Piotr D Adamczyk and Brian P Bailey. 2004. If not now, when? The effects of interruption at different moments within task execution. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 271–278. doi:10.1145/985692.985700

[2] Taslima Akter, Yoonha Cha, Isabela Figueira, Stacy M. Branham, and Anne Marie Piper. 2023. "If I'm supposed to be the facilitator, I should be the host": Understanding the Accessibility of Videoconferencing for Blind and Low Vision Meeting Facilitators. In *Proceedings of the 25th International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS '23)*. doi:10.1145/3597638.3608420

[3] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millicah, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. 2022. Flamingo: a visual language model for few-shot learning. In *Proceedings of the 36th International Conference on Neural Information Processing Systems* (New Orleans, LA, USA) *(NIPS '22)*. Curran Associates Inc., Red Hook, NY, USA, Article 1723, 21 pages.

[4] CAST. 2018. Universal Design for Learning guidelines version 2.2. https://udlguidelines.cast.org

[5] Xinyue Chen, Kunlin Ruan, Kexin Phyllis Ju, Nathan Yap, and Xu Wang. 2025. More AI Assistance Reduces Cognitive Engagement: Examining the AI Assistance Dilemma in AI-Supported Note-Taking. *Proc. ACM Hum.-Comput. Interact.* 9, 7, Article CSCW451 (Oct. 2025), 29 pages. doi:10.1145/3757632

[6] Gheorghe Comanici et al. 2025. Gemini 2.5: Pushing the Frontier. https://arxiv.org/abs/2507.06261

[7] Maitraye Das, John Tang, Kathryn E. Ringland, and Anne Marie Piper. 2021. Towards Accessible Remote Work: Understanding Work-from-Home Practices of Neurodivergent Professionals. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW1 (2021). doi:10.1145/3449282

[8] Google Gemini Team. 2024. Gemini 2.5: A Family of Highly Capable Multimodal Models. https://arxiv.org/abs/2312.11805

[9] Raja S. Kushalnagar et al. 2014. Captions or Transcripts? A Comparison of Accessibility Tools. In *Proceedings of the 16th International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS '14)*. doi:10.1145/2661334.2661381

[10] Anne-Laure Le Cunff, Vincent Giampietro, and Eleanor Dommett. 2024. Neurodiversity and cognitive load in online learning: A systematic review with narrative synthesis. *Educational Research Review* 43 (2024), 100604. doi:10.1016/j.edurev.2024.100604

[11] Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. 2024. LLaVA-NeXT-Interleave: Tackling Multi-image, Video, and 3D in Large Multimodal Models. arXiv:2407.07895 [cs.CV] https://arxiv.org/abs/2407.07895

[12] Richard E Mayer. 2009. *Multimedia Learning* (2 ed.). Cambridge University Press. doi:10.1017/CBO9780511811678

[13] Richard E Mayer and Roxana Moreno. 2003. Nine ways to reduce cognitive load in multimedia learning. *Educational Psychologist* 38, 1 (2003), 43–52. doi:10.1207/S15326985EP3801_6

[14] Cecily Morrison, Rita Faia Marques, Martin Grayson, Daniela Massiceti, Camilla Longden, Linda Yilin Wen, and Ed Cutrell. 2023. Understanding Personalized Accessibility through Teachable AI: Designing and Evaluating Find My Things for People who are Blind or Low Vision. In *ASSETS 2023*. 1–12. https://www.microsoft.com/en-us/research/publication/understanding-personalized-accessibility-through-teachable-ai-designing-and-evaluating-find-my-things-for-people-who-are-blind-or-low-vision/

[15] Amy Pavel et al. 2024. Describe Now: User-Driven Audio Description for Blind and Low Vision Individuals. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. doi:10.1145/3715336.3735685

[16] Kilian G. Seeber. 2011. Cognitive load in simultaneous interpreting: Existing theories — new models. *Interpreting* 13, 2 (2011), 176–204. doi:10.1075/intp.13.2.02see

[17] Abigale Stangl et al. 2024. The kinds of description BLV people need from different kinds of videos. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. doi:10.1145/3613904.3642238

[18] Josh Starmer. 2025. Reinforcement Learning with Human Feedback (RLHF), Clearly Explained!!! YouTube. https://www.youtube.com/watch?v=qPN_XZcJf_s StatQuest with Josh Starmer.

[19] John Sweller. 1988. Cognitive load during problem solving: Effects on learning. *Cognitive Science* 12, 2 (1988), 257–285. doi:10.1016/0364-0213(88)90023-7

[20] Mike Wald. 2021. AI data-driven personalisation and disability inclusion. *Frontiers in artificial intelligence* 3 (2021), 571955.

[21] Lu Wang, Yilong Li, Jianhua He, Yueling Che, Kaishun Wu, Xiaoke Qi, and Kaixin Chen. 2026. MeetSumAid: A Mobile Human-AI Collaborative Meeting Summarization System. *IEEE Transactions on Mobile Computing* (2026), 1–16. doi:10.1109/TMC.2026.3667272

[22] Anna Wong, Wayne Leahy, Nadine Marcus, and John Sweller. 2012. Cognitive load theory, the transient information effect and e-learning. *Learning and Instruction* 22, 6 (2012), 449–457. doi:10.1016/j.learninstruc.2012.05.004

[23] Farnaz Zamiri Zeraati, Yang Trista Cao, Yuehan Qiao, Hal Daumé III, and Hernisa Kacorri. 2026. Say It My Way: Exploring Control in Conversational Visual Question Answering with Blind Users. *arXiv preprint arXiv:2602.16930* (2026).

[24] Ling Zhang, Richard Allen Carter Jr, Matthew L Bernacki, and Jeffrey A Greene. 2025. Personalization, individualization, and differentiation: What do they mean and how do they differ for students with disabilities? *Exceptionality* 33, 4 (2025), 241–262.