International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 11, CONFLUENCE - 2026

An Intelligent Multimodal System for Continuous Lecture Note Generation Using Voice and Visual Content

Thimmaiah K K S[1], Student, Bachelor of Computer Applications (BCA)
St. Francis Degree College, Koramangala, Bengaluru Affiliated to Bengaluru City University, Karnataka, India
Jyothiba Salunkhe S[2] , Assistant Professor, Bachelor of Computer Applications (BCA)
St. Francis Degree College, Koramangala, Bengaluru Affiliated to Bengaluru City University, Karnataka, India, **salunkhejyothiba@gmail.com**

Abstract - In this paper, I describe VoxBox, an Android application intended to capture both sources of lecture content and produce structured study notes while retaining the original evidence for later review. A student follows spoken explanations and material on the board or screen, including slides, equations and diagrams, at the same time. Voice recorders and transcription tools preserve speech without the board; photographs preserve the board without the explanation that gives it meaning. Voice mode serves discussion-led sessions, whereas Live Board mode adds periodic CameraX frames to continuous audio capture. Recorded in bounded 20-second chunks, audio remains on the device until transcription succeeds. Before any frame leaves the phone, a local visual gate and on-device optical character recognition classify it as a duplicate, a continuation or a new scene. The remote extraction service receives only finalised board segments. Transcripts and board images are stored separately from the AI-generated notes and are never rewritten. For note generation, the system uses revision-guarded Markdown, bounded syllabus context and an authenticated backend proxy, with provenance linking each statement to its source. A survey of sixteen recent studies (2022–2025) highlights the gaps targeted by the design. In a two-minute device session, 18 frames were captured, of which 15 were rejected as duplicates and 2 merged as continuations. This left one representative image and no board extraction request. All 193 automated tests pass in the current build. Classroom-scale evaluation of accuracy and endurance remains future work.

Keywords: Multimodal lecture notes; automatic speech recognition; optical character recognition; evidence preservation; responsible educational technology

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 11, CONFLUENCE - 2026

1. INTRODUCTION

In a lecture, reasoning and worked explanations come through speech, alongside definitions, equations and diagrams on the board or slide. The student follows both at once, so taking handwritten notes on the two sources divides attention. A recent meta-analysis suggests that handwritten notes are linked to higher achievement, although typing captures more content (Flanigan et al., 2024). Existing digital aids appear to cover only one source: voice recorders and transcription services omit the board, while photographs preserve it without the explanation that gives it meaning.

Evidence preservation is the defining principle of VoxBox, the Android application I present here. Its raw transcript and board images are stored before any AI interpretation; those records are never rewritten and remain available whenever a generated statement is in doubt. The application generates structured study notes from spoken explanations and board or screen content during a lecture. The paper makes three contributions:

- A structured survey of sixteen recent studies (2022–2025) on lecture analysis, speech recognition, optical character recognition (OCR) and grounded text generation (Section 2);
- An architecture in which evidence is preserved and an on-device board-filtering pipeline classifies every frame before any remote request is made (Section 4); and
- Device-level results reported with explicit claim boundaries, together with an evaluation protocol for the measurements still outstanding (Sections 5–7).

2. LITERATURE SURVEY

A multimodal lecture-note system draws on four areas of research: (A) Note-taking and lecture understanding, (B) Speech recognition, (C) Visual text and mathematics recognition, and (D) Grounded and verifiable text generation. This survey covers all four areas through sixteen studies published between 2022 and 2025. Each study was selected for its relevance to one stage of the VoxBox pipeline. Table 1 presents their methods and the gaps most relevant to continuous lecture note generation.

Table 1. Literature survey of related work

Sl. No.	Paper Name	Authors	Method Used	Key Gaps & Limitations
A. Note-taking and lecture understanding				
1	Typed Versus Handwritten Lecture Notes and College Student	Flanigan, A. E., Wheeler, J., Colliot, T., et al. (2024)	A meta-analysis of 24 studies from 21 articles compares typed and	Handwritten notes were linked to higher achievement (g = 0.248), although typing produced

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 11, CONFLUENCE - 2026

Sl. No.	Paper Name	Authors	Method Used	Key Gaps & Limitations
	Achievement: A Meta-Analysis *Educational Psychology Review*		handwritten lecture notes among college students.	more notes (g = 0.919). The analysis covers manual note-taking only and does not study AI-generated notes.
2	Lecture Presentations Multimodal Dataset: Towards Understanding Multimodality in Educational Videos *IEEE/CVF ICCV*	Lee, D. W., Ahuja, C., Liang, P. P., et al. (2023)	The dataset contains 180+ hours of lecture video and 9,000+ slides aligned with speech. The work benchmarks cross-modal retrieval and proposes the PolyViLT transformer.	The authors report weak slide–speech alignment, novel visual media, technical language and long sequences as open challenges. The recordings show slides rather than live boards.
3	M3AV: A Multimodal, Multigenre, and Multipurpose Audio-Visual Academic Lecture Dataset *ACL*	Chen, Z., Liu, H., Yu, W., et al. (2024)	About 367 hours of academic lecture video are accompanied by human-annotated slide text and speech. Benchmarks cover contextual ASR, speech synthesis, and slide and script generation.	The work provides a dataset and benchmark for recorded presentations with projected slides rather than handwritten classroom boards. It is not an end-to-end note system.
4	NoTeeline: Supporting Real-Time, Personalized Notetaking with LLM-Enhanced Micronotes *ACM IUI*	Huq, F., Samee, A., Lin, D. C.-E., et al. (2025)	As a video plays, an LLM expands short user-written micronotes into full notes in the user's style. The system was evaluated in a within-subjects study (n = 12).	Factual correctness was 93.2%, so errors still need to be identified. The sample was small and the videos pre-recorded. Students must continue writing micronotes during the lecture.
B. Speech recognition				
5	Robust Speech Recognition via Large-Scale Weak Supervision *ICML*	Radford, A., Kim, J. W., Xu, T., et al. (2023)	Whisper is an encoder–decoder Transformer trained on 680,000 hours of weakly supervised audio. It decodes long recordings sequentially in 30 s windows.	For long-form audio, the authors report repeat loops, missed words at segment edges and hallucinated text. The system processes audio only and provides no speaker attribution.
6	WhisperX: Time-Accurate Speech Transcription of Long-Form Audio *Interspeech*	Bain, M., Huh, J., Han, T., & Zisserman, A. (2023)	WhisperX combines voice-activity cut-and-merge into ~30 s chunks, batched Whisper inference and forced phoneme alignment for word-level timestamps.	WhisperX requires additional VAD and language-specific alignment models. It uses batch GPU processing of complete recordings rather than live mobile capture.
7	Turning Whisper into Real-Time Transcription System *IJCNLP-AACL System Demonstrations*	Macháček, D., Dabre, R., & Bojar, O. (2023)	A local-agreement policy with self-adaptive latency adapts Whisper for streaming transcription. Latency is about 3.3 s on long-form speech. The transcriber was demonstrated at a live conference.	The transcriber optimises the latency of a rolling transcript without adding board content, note structure or speaker identity.

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 11, CONFLUENCE - 2026

Sl. No.	Paper Name	Authors	Method Used	Key Gaps & Limitations
8	Careless Whisper: Speech-to-Text Hallucination Harms *ACM FAccT*	Koenecke, A., Choi, A. S. G., Mei, K. X., et al. (2024)	An audit of Whisper output on 13,140 AphasiaBank segments identifies hallucinations and categorises them by type of harm.	1.4% of transcriptions contained hallucinations, 38% of them harmful, more often with long pauses. The work diagnoses the problem without a mitigation pipeline.
9	SlideSpeech: A Large Scale Slide-Enriched Audio-Visual Corpus *IEEE ICASSP*	Wang, H., Yu, F., Shi, X., et al. (2024)	SlideSpeech contains 1,705 conference videos (1,000+ hours, 473 transcribed) with synchronised slides. Slide keywords provide context for speech recognition.	Projected slides from recorded talks supply text only as recognition context. Handwritten boards and note generation are not included.
C. Visual text and mathematics recognition				
10	TrOCR: Transformer-Based Optical Character Recognition with Pre-trained Models *AAAI*	Li, M., Lv, T., Chen, J., et al. (2023)	TrOCR's pre-trained image-Transformer encoder and text-Transformer decoder are fine-tuned for printed, handwritten and scene text.	A separate detector is needed because recognition operates on cropped text lines. Continuous use on a phone is costly, and there is no layout or equation structure.
11	CoMER: Modeling Coverage for Transformer-Based Handwritten Mathematical Expression Recognition *ECCV*	Zhao, W., & Gao, L. (2022)	A Transformer decoder recognises handwritten mathematical expressions using coverage attention (attention refinement, self- and cross-coverage).	Expression recognition rate is about 59–63% on CROHME benchmarks. The method recognises isolated expressions rather than full camera-captured boards.
12	Nougat: Neural Optical Understanding for Academic Documents *ICLR*	Blecher, L., Cucurull Preixens, G., Scialom, T., & Stojnic, R. (2024)	A Visual Transformer converts images of scientific document pages into markup, including mathematical expressions.	The model is designed for digital academic documents and does not address handwriting, camera perspective or glare on classroom boards.
13	OCRBench: On the Hidden Mystery of OCR in Large Multimodal Models *Science China Information Sciences*	Liu, Y., Li, Z., Huang, M., et al. (2024)	A benchmark of 29 datasets evaluates large multimodal models such as GPT-4V and Gemini. It covers text recognition, VQA, information extraction and handwritten mathematics.	Models show clear weaknesses with handwritten text, non-semantic text and mathematical expressions, all typical board content.
D. Grounded and verifiable text generation				
14	FActScore: Fine-grained Atomic Evaluation of Factual Precision in Long Form Text Generation *EMNLP*	Min, S., Krishna, K., Lyu, X., et al. (2023)	FActScore splits generated text into atomic facts and measures the share supported by a knowledge source through human and automated evaluation.	On biographies, ChatGPT reached only 58%. Facts are checked against an external source rather than the input used to generate a note.

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 11, CONFLUENCE - 2026

Sl. No.	Paper Name	Authors	Method Used	Key Gaps & Limitations
15	Enabling Large Language Models to Generate Text with Citations *EMNLP*	Gao, T., Yen, H., Yu, J., & Chen, D. (2023)	In ALCE, LLM systems retrieve evidence and answer with citations. The benchmark scores fluency, correctness and citation quality.	On ELI5, even the best models lacked complete citation support about half the time. The task is question answering rather than lecture notes.
16	Retrieval-Augmented Generation for Large Language Models: A Survey *arXiv preprint*	Gao, Y., Xiong, Y., Gao, X., et al. (2024)	The survey covers naive, advanced and modular RAG, including retrieval, generation, augmentation and evaluation methods.	Retrieval mitigates hallucination but does not eliminate it. Retrieved text is reference knowledge, not evidence of what a lecturer said.

2.1 Synthesis of the Survey

In Table 1, I identify three patterns. Recent lecture datasets and note-taking tools (Chen et al., 2024; Huq et al., 2025; Lee et al., 2023) work with recorded presentations and projected slides. None of them appears to capture a live classroom from a student's phone, where speech and a handwritten board change continuously. The components are strong but fallible: recognisers hallucinate during pauses and miss words at segment edges (Koenecke et al., 2024; Radford et al., 2023). Streaming calls for careful boundary handling (Macháček et al., 2023); slide text can aid recognition (Wang et al., 2024). Both specialised recognisers and multimodal models remain weak on handwriting and mathematics (Liu et al., 2024; Zhao & Gao, 2022). LLM output contains unsupported statements even when it cites sources (T. Gao et al., 2023; Min et al., 2023). Retrieved context improves factuality without establishing what was actually said (Y. Gao et al., 2024). I take these findings to mean that no single stage should be trusted on its own. Evidence should therefore be retained between stages, deterministic work run locally and remote output kept traceable to its evidence.

3. RESEARCH GAP AND OBJECTIVES

Transcript-centred tools miss board equations and diagrams. Camera-centred tools miss the reasoning spoken while the board changes. Summarisation-centred tools obscure whether a statement came from speech, the board or the model. None of the surveyed systems appears to combine live, two-channel capture on a mobile device with an auditable separation between evidence and generated notes. These gaps lead me to four questions:

RQ1 How can a phone capture speech and board evidence together without losing the final partial speech segment or sending redundant frames?

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 11, CONFLUENCE - 2026

RQ2 How can notes be restructured freely while the underlying evidence remains intact and traceable?

RQ3 Which privacy, provenance and recovery controls must be in place before AI-generated notes can be trusted?

RQ4 What must be measured before claims can be made about accuracy, reliability or cost?

The objective is an Android lecture companion whose output is portable Markdown notes backed by auditable evidence. The architecture answers RQ1–RQ3; an evaluation protocol answers RQ4.

4. METHODOLOGY AND SYSTEM ARCHITECTURE

The work follows a design-science method: problem identification, technology review, artefact construction, device verification and an evaluation protocol for claims not yet measured. I present the architecture in Figure 1. A session starts when the user selects a subject, a mode (Voice or Live Board) and a note policy, and capture runs only while the application is in the foreground. Independent workers handle audio, local frames and remote extraction. Camera load therefore cannot delay ordered speech. Durable state is stored on the phone, and a hosted proxy mediates every model call. The provider credentials are held only by the proxy.

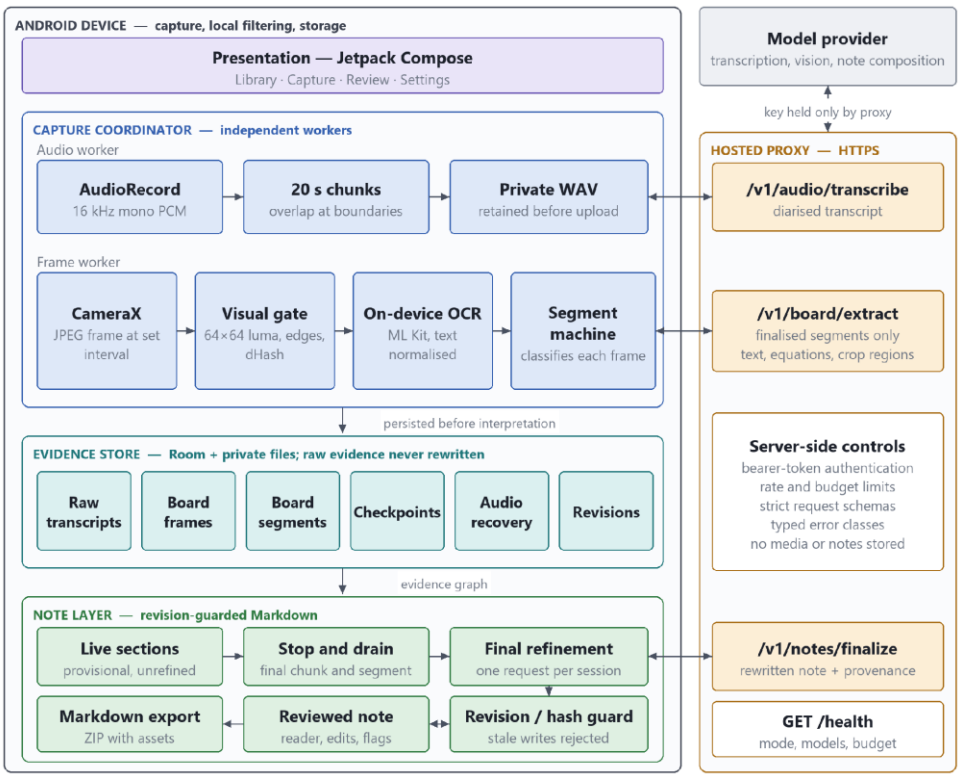


Figure 1. System architecture of VoxBox: independent capture workers, a local evidence store, a revision-guarded note layer and an authenticated proxy boundary

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 11, CONFLUENCE - 2026

4.1 Audio Evidence Pipeline

Android's AudioRecord API records audio as 16 kHz mono PCM in 20-second chunks. The platform SpeechRecognizer was rejected on the basis of Android's own documentation, which advises against using it for continuous recognition on battery and bandwidth grounds. Before upload, each chunk is saved privately as a WAV file. Consecutive chunks overlap slightly so that a word falling on a boundary is less likely to be lost (Macháček et al., 2023; Radford et al., 2023). The final partial chunk is drained on stop rather than discarded. The on-device Room database stores transcripts as immutable, timestamped segments. Only after its transcript has been persisted is a WAV file deleted. A transcription failure leaves the file available for recovery. Assigned per chunk, speaker labels serve only as a local hint about the dominant speaker, never as an identity.

4.2 Live Board Filtering Pipeline

Live Board mode adds periodic frame capture through the CameraX library. Boards rarely change all at once; most updates add a single word or equation line. Sending every frame to a vision model would therefore be wasteful. Below, I describe the three on-device stages each frame passes through first:

(a) **Visual gate.** The gate compares a 64 × 64 luminance sample with the baseline using exposure-compensated novelty, changed-pixel fraction, edge statistics and a 64-bit difference hash. It rejects flicker, sensor noise and pointer movement without OCR.

(b) **On-device OCR.** The bundled ML Kit text recogniser reads the surviving frames, which are then compared with the open segment by token, edit-distance and line-containment similarity.

(c) **Segment state machine.** Each frame is classified as shown in Table 2. Only a finalised segment, represented by its best frame, is sent for remote extraction of text, equations and diagram regions.

Table 2. Frame classification and local action

Classification	Typical board event	Local action
Duplicate	Unchanged board, exposure flicker or a small pointer movement	Discard the redundant image after persistence and keep the clearer representative; no remote request.
Continuation	A word, bullet or equation line is added	Merge the new text into the open segment; defer the remote request.
New scene	Slide change or erased board	Finalise the open segment for extraction and start a new one.

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 11, CONFLUENCE - 2026

Classification	Typical board event	Local action
Non-text change	A diagram grows without new text	Keep the frame in the open segment as a diagram candidate.
Uncertain	Large visual change with little readable text	Hold one further frame; unresolved frames are retained for review.

The heuristic score used to select the representative frame considers sharpness, exposure, obstruction, completeness, OCR length and diagram visibility. A segment is flushed on a new scene, on stop or manual capture, or when it reaches its maximum age (180 s by default) or capacity. Recovery after process death also triggers a segment flush. Checkpointing allows an interrupted session to resume without replaying frames.

4.3 Note Generation and Provenance

Users can select one of three note policies. *Verbatim* formats timestamped speech and board text without AI rewriting. *Evidence strict* rewrites conservatively, adds no claims from the syllabus or general knowledge and flags uncertain gaps. *Study enhanced*, the default policy, is instructed to merge repeated explanations, reconcile speech with board text, repair fragments and remove unrelated conversation. Every change is recorded as provenance.

While capture is in progress, live sections are composed locally and marked as unrefined. After stop-and-drain, a single finalisation request sends the complete evidence graph (transcripts, board segments, ranked syllabus excerpts and the existing note) to the proxy. The proxy returns a rewritten note with each section labelled as captured, normalised, corrected or AI-supplemented. Correction decisions give transcript–board agreement priority over repeated lecture context and the syllabus. Each correction records the original and proposed wording, a reason, a confidence value and its evidence. Contradictions are flagged for review. Syllabus text is never presented as something the lecturer said. Every write, whether from the model or the user, commits only if its base revision and hash match the current note. Stale responses are rejected.

5. IMPLEMENTATION AND RESULTS

The re-architected build was verified on a Redmi 2411DRN47I running Android 16 (Figure 2), using a loopback mock proxy with the provider transport disabled. No billable request was made. These runs therefore do not establish provider accuracy or note quality. Table 3 lists each result and the claims it does not establish.

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 11, CONFLUENCE - 2026

Table 3. Implementation results and claim boundaries

Area	Observed result	Not established
Automated tests	115/115 Android unit, 66/66 backend and 12/12 on-device instrumentation tests pass (193/193).	Classroom accuracy.
Audio capture	20 s chunks transcribed in order; final partial chunk drained at stop; recovery files cleared after success.	Word error rate (transcription used the mock).
Board filtering (device)	Two-minute session: 18 frames captured, 15 duplicates skipped, 2 continuations merged, 0 extraction requests, 1 representative image kept.	Performance on changing boards; the scene was largely static.
Board filtering (replay)	30 synthetic captures over 3 topics: 9 local OCR passes, 3 segments; 6 of 9 extraction requests avoided (66.67%).	Monetary cost or real error rates.
Camera zoom	3.53× held across 18 captures and 75 s; an automated test held 2.0× for 60 s and across activity recreation.	Behaviour on other phone models.
Export	ZIP of UTF-8 Markdown with YAML front matter, relative asset paths and separate evidence sections.	Import into an external vault (not re-tested).

Repetition removal, gap filling and correction remain unverified because the mock returns one fixed sample per audio chunk. In field trials, an earlier prototype connected to a live provider transcribed real speech with timestamps and speaker labels. This suggests that the approach is feasible, but its accuracy has yet to be established.

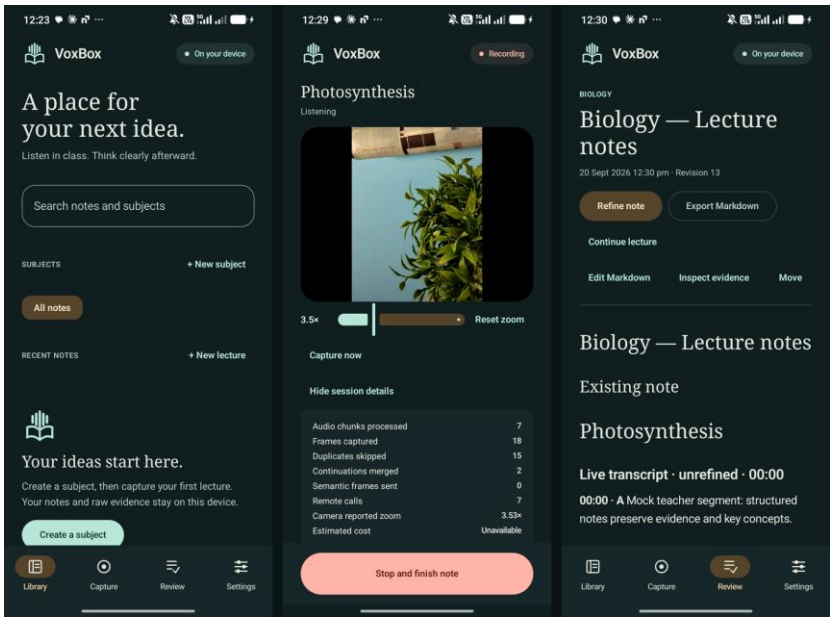


Figure 2. VoxBox on a physical device (Android 16): the library, a Live Board session with its diagnostics panel, and the reviewed note

6. PRIVACY, SECURITY AND RESPONSIBLE USE

The user starts and stops each session. Capture runs only in the foreground, and the application requests only the permissions needed for the selected mode when capture begins. Users are responsible for obtaining consent before recording a class. Provider credentials never enter the

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 11, CONFLUENCE - 2026

application. The HTTPS proxy requires a bearer token, enforces size, rate and budget limits, rejects requests that do not conform to a strict schema, and does not write submitted media, syllabi or notes to disk. The client token can be extracted from the APK, so its risk is bounded by those limits rather than by secrecy. Speaker labels are never used to identify people, as identification would require consent-aware enrolment. Generated notes are presented as revision aids, with their source evidence available for inspection.

7. EVALUATION PLAN

Accuracy claims require a labelled corpus fixed before any tuning (Table 4). Each run will record the device, build, model configuration and network condition. Results will include sample sizes and uncertainty. The evaluation will compare voice-only and multimodal capture to measure concept coverage. Uploading every changed frame will be compared with local filtering to measure requests avoided versus changes missed. Direct summarisation will be compared with evidence-linked generation for unsupported statements and review effort.

Table 4. Planned evaluation dimensions and metrics

Dimension	Test material	Metrics
Speech	Quiet and noisy rooms, competing speakers, long silences, short final chunks	Word error rate, timestamp error, diarisation error rate, final-chunk preservation
Board	Static frames, progressive reveals, equations, diagrams, exposure shifts, erasures, slide changes	Duplicate-rejection precision, new-scene recall, OCR and equation accuracy, diagram-region IoU, requests avoided
Note quality	Reference concept lists linked to evidence identifiers	Concept coverage, duplicate-concept rate, unsupported-statement rate, correction precision
Resources	30- to 60-minute sessions on stated hardware	Latency, network requests, storage growth, battery and thermal behaviour
Usability	Task-based study with students	Task completion, correction time, accessibility (large text, TalkBack), export validity

8. DISCUSSION AND LIMITATIONS

VoxBox's main strength is that it retains the evidence alongside the generated notes. A statement in doubt can be traced to its transcript segment or board frame, and media from a failed request remain available for reprocessing. Local filtering should make session cost depend on the number of distinct board states rather than lecture length, although this has yet to be measured. The main limitation is that accuracy remains unmeasured. Each component has documented failure modes (Table 1), and verification has so far used a mock provider on a single phone model. The design also has limits: OCR covers Latin script only, syllabus import

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 11, CONFLUENCE - 2026

accepts only UTF-8 text and Markdown, diagrams are cropped without interpretation, and very long lectures are rejected rather than consolidated. No 30-minute endurance, battery, thermal or process-death test has yet been run. Results from one classroom, subject or device may not generalise.

9. CONCLUSION AND FUTURE WORK

VoxBox generates continuous lecture notes from voice and visual content and keeps the evidence those notes derive from. The survey suggests that the required components exist but are individually fallible, and that recent lecture systems work with recorded presentations rather than live classrooms. VoxBox responds with visible capture, recoverable audio chunks, on-device board filtering, immutable evidence, revision-guarded generation and a credential-isolating proxy. The build passes all 193 automated tests. In one device session, redundant board frames were eliminated before any remote request was made.

Future work will proceed in order of evidential value: a live-provider session with speech and a changing board comes first, then a labelled corpus and scorer for Table 4, then endurance and recovery testing. Handwriting and equation recognition, long-lecture consolidation and consent-aware speaker identification come last. The system should claim only what its evidence supports.

REFERENCES

[1] **Bain, M., Huh, J., Han, T., & Zisserman, A. (2023).** WhisperX: Time-accurate speech transcription of long-form audio. In *Proceedings of Interspeech 2023* (pp. 4489–4493). International Speech Communication Association. https://doi.org/10.21437/Interspeech.2023-78

[2] **Blecher, L., Cucurull Preixens, G., Scialom, T., & Stojnic, R. (2024).** Nougat: Neural optical understanding for academic documents. In *Proceedings of the Twelfth International Conference on Learning Representations (ICLR 2024)*. https://proceedings.iclr.cc/paper_files/paper/2024/hash/a39a9aceda771cded859ae7560530e09-Abstract-Conference.html

[3] **Chen, Z., Liu, H., Yu, W., Sun, G., Liu, H., Wu, J., Zhang, C., Wang, Y., & Wang, Y. (2024).** M3AV: A multimodal, multigenre, and multipurpose audio-visual academic lecture dataset. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 9041–9060). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.acl-long.489

[4] **Flanigan, A. E., Wheeler, J., Colliot, T., Lu, J., & Kiewra, K. A. (2024).** Typed versus handwritten lecture notes and college student achievement: A meta-analysis. *Educational Psychology Review, 36*(3), Article 78. https://doi.org/10.1007/s10648-024-09914-w

[5] **Gao, T., Yen, H., Yu, J., & Chen, D. (2023).** Enabling large language models to generate text with citations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (pp. 6465–6488). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.emnlp-main.398

[6] **Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., Wang, M., & Wang, H. (2024).** *Retrieval-augmented generation for large language models: A survey* (arXiv:2312.10997). arXiv. https://doi.org/10.48550/arXiv.2312.10997

[7] **Huq, F., Samee, A., Lin, D. C.-E., Tang, A. X., & Bigham, J. P. (2025).** NoTeeline: Supporting real-time, personalized notetaking with LLM-enhanced micronotes. In *Proceedings of the 30th International Conference on Intelligent User Interfaces* (pp. 1064–1081). Association for Computing Machinery. https://doi.org/10.1145/3708359.3712086

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 11, CONFLUENCE - 2026

[8] **Koenecke, A., Choi, A. S. G., Mei, K. X., Schellmann, H., & Sloane, M. (2024).** Careless Whisper: Speech-to-text hallucination harms. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency* (pp. 1672–1681). Association for Computing Machinery. https://doi.org/10.1145/3630106.3658996

[9] Lee, D. W., Ahuja, C., Liang, P. P., Natu, S., & Morency, L.-P. (2023). Lecture presentations multimodal dataset: Towards understanding multimodality in educational videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)* (pp. 20030–20041). IEEE. https://doi.org/10.1109/ICCV51070.2023.01838

[10] **Li, M., Lv, T., Chen, J., Cui, L., Lu, Y., Florencio, D., Zhang, C., Li, Z., & Wei, F. (2023).** TrOCR: Transformer-based optical character recognition with pre-trained models. *Proceedings of the AAAI Conference on Artificial Intelligence*, *37*(11), 13094–13102. https://doi.org/10.1609/aaai.v37i11.26538

[11] **Liu, Y., Li, Z., Huang, M., Yang, B., Yu, W., Li, C., Yin, X.-C., Liu, C.-L., Jin, L., & Bai, X. (2024).** OCRBench: On the hidden mystery of OCR in large multimodal models. *Science China Information Sciences*, *67*(12), Article 220102. https://doi.org/10.1007/s11432-024-4235-6

[12] **Macháček, D., Dabre, R., & Bojar, O. (2023).** Turning Whisper into real-time transcription system. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics: System Demonstrations* (pp. 17–24). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.ijcnlp-demo.3

[13] **Min, S., Krishna, K., Lyu, X., Lewis, M., Yih, W.-t., Koh, P. W., Iyyer, M., Zettlemoyer, L., & Hajishirzi, H. (2023).** FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (pp. 12076–12100). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.emnlp-main.741

[14] **Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2023).** Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning* (Vol. 202, pp. 28492–28518). PMLR. https://proceedings.mlr.press/v202/radford23a.html

[15] **Wang, H., Yu, F., Shi, X., Wang, Y., Zhang, S., & Li, M. (2024).** SlideSpeech: A large scale slide-enriched audio-visual corpus. In *ICASSP 2024 – 2024 IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 11076–11080). IEEE. https://doi.org/10.1109/ICASSP48485.2024.10448079

[16] **Zhao, W., & Gao, L. (2022).** CoMER: Modeling coverage for transformer-based handwritten mathematical expression recognition. In S. Avidan, G. Brostow, M. Cissé, G. M. Farinella, & T. Hassner (Eds.), *Computer vision – ECCV 2022* (Lecture Notes in Computer Science, Vol. 13688, pp. 392–408). Springer. https://doi.org/10.1007/978-3-031-19815-1_23

(This work is licensed under a Creative Commons Attribution 4.0 International License.)