IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 2, October 2026
Impact Factor: 8.2

AI-Powered Multimodal Content Transformation Platform

Lavanya Lalit Dixit, Neha Kailas Ahirrao, Vaishnavi Atul Thoke,
Harshada Sunil Pawar, Prof. O. S. Korde
Department of Computer Engineering
Guru Gobind Singh Polytechnic, Nashik

Abstract: *Digital information is distributed across images, text, audio, and video, while conversion between these formats often requires multiple separate tools. This literature review examines research relevant to an AI-powered multimodal content transformation platform that brings seven operations into one web-based workflow: image-to-text, text-to-image, text-to-audio, audio-to-text, video-to-audio, video-to-text, and image-to-audio. Prior work on multimodal deep learning provides the broader foundation for connecting information across modalities [1]. Research on automatic speech recognition discusses deep learning approaches and challenges such as domain variation and computational requirements [2]. Generative models support text-conditioned image synthesis [3], while deep-learning-based speech synthesis enables the conversion of text into spoken output [4]. Video-captioning research combines computer vision and natural-language generation to describe video content [5]. Privacy-aware distributed learning is also relevant when user content is sensitive [6]. The reviewed literature supports the feasibility of the platform's individual conversion modules, but does not establish that all seven operations are already implemented or validated in a single system. The review identifies integration, quality evaluation, accessibility, privacy, and reliable output generation as key considerations..*

Keywords: Multimodal AI, content transformation, OCR, speech recognition, text-to-speech, text-to-image, video captioning, generative AI.

I. INTRODUCTION

Digital content commonly appears as scanned notes, photographs, written documents, speech recordings, and videos. Users who need to reuse such content may have to move between separate applications for optical character recognition (OCR), speech recognition, text-to-speech (TTS), image generation, and video processing. This fragmented workflow can increase the time required to complete simple tasks and create additional steps for students, teachers, content creators, and people who benefit from accessible formats.

The supplied project presentation proposes a unified web-based platform that applies artificial intelligence and machine learning to transform content between modalities. Its seven listed operations are image-to-text, text-to-image, text-to-audio, audio-to-text, video-to-audio, video-to-text, and image-to-audio. The goal is to provide a simple workflow in which users select an operation, supply or upload input, and receive a generated result. The platform is intended to support education, content creation, accessibility, digital documentation, and general multimedia use.

II. REVIEW SCOPE

This is a focused narrative review organized around the conversion operations named in the project presentation. It draws on selected review papers and research articles about multimodal deep learning, automatic speech recognition, text-to-image generation, expressive speech synthesis, video captioning, and multimodal federated learning. The discussion summarizes each source's relevance to the proposed system and identifies practical integration issues. It is not a systematic review, meta-analysis, or experimental assessment of the proposed platform. The presentation does not

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 2, October 2026
Impact Factor: 8.2

provide implementation measurements, model names for every module, or benchmark results; therefore, no performance claims are inferred.

III. LITERATURE REVIEW

A. Multimodal Deep Learning

Jabeen et al. review methods and applications in multimodal deep learning, a field concerned with learning from and connecting different forms of data [1]. This provides the broad conceptual basis for a platform that handles text, images, audio, and video through one interface. For the proposed system, multimodal thinking helps frame conversion as a coordinated set of tasks rather than unrelated tools. However, a survey of multimodal methods does not itself specify the architecture or prove the reliability of a particular seven-operation application.

B. Audio-to-Text: Automatic Speech Recognition

Kheddar, Hemis, and Himeur survey advanced deep-learning approaches for automatic speech recognition (ASR) [2]. They discuss approaches including transformers, transfer learning, federated learning, and reinforcement learning, as well as challenges related to domain differences and resource demands. This work is directly relevant to the audio-to-text module, which needs to turn spoken input into readable text. In a practical platform, recognition quality may depend on recording quality, accents, background noise, language, and the chosen model. Such factors should be evaluated rather than assuming that transcription is always exact.

C. Text-to-Image: Generative Models

Research on text-to-image synthesis studies how generative models create visual content from natural-language prompts [3]. This research area underpins the text-to-image operation in the proposed platform. The user can provide a description and receive a generated image, potentially supporting learning materials and content creation. Important considerations include prompt adherence, image quality, generation time, computational cost, and the possibility that the generated image may not accurately represent all details requested by the user. The supplied presentation names generative AI as a core technology but does not identify a specific image-generation model.

D. Text-to-Audio: Speech Synthesis

Barakat, Turk, and Demiroglu provide a systematic review of deep-learning-based expressive speech synthesis [4]. Their review notes that modern TTS systems aim not only for intelligibility but also for expressive delivery and varied speaking styles. This literature supports the text-to-audio module, which can make written material available as speech. Potential applications include listening to study notes and providing audio alternatives to text. Output should be assessed for intelligibility, naturalness, pronunciation, language coverage, and suitability for the intended use.

E. Video-to-Text: Video Captioning

Vaishnavi and Narmatha survey video captioning, a task that combines visual analysis and natural-language generation to describe events and content in video [5]. This is relevant to the platform's video-to-text operation, where a user expects a textual description or summary of video content. Video captioning differs from single-image description because the system may need to identify actions and events over time. A useful implementation should distinguish between visible evidence and inferred context, and its output should be checked for omitted events, incorrect descriptions, and temporal misunderstandings.

F. Privacy and Multimodal Learning

Che et al. survey multimodal federated learning, which explores collaborative learning across distributed data sources while addressing privacy concerns [6]. Although federated learning is not stated as an implemented feature in the project presentation, this work is relevant to future development where uploaded images, recordings, or documents may

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 2, October 2026
Impact Factor: 8.2

contain personal or confidential information. A web-based platform should clearly explain how inputs are processed and stored, minimize unnecessary retention, and use appropriate access controls. Privacy-preserving training is a possible research direction rather than a confirmed feature of the current project.

G. OCR and Image-Based Accessibility

The project presentation identifies OCR as a core technology for image-to-text and image-to-audio conversion. OCR extracts machine-readable text from images or scanned pages; the extracted text can then be displayed or passed to a speech-synthesis module. This creates a useful accessibility pipeline for image-based study notes and documents. The presentation does not name a specific OCR engine or provide accuracy results, so recognition quality for different fonts, layouts, image resolutions, and languages remains to be tested.

IV. COMPARATIVE ANALYSIS

Table I maps each reviewed source to the part of the proposed platform it informs, together with the main considerations that follow from it.

TABLE I: SELECTED LITERATURE MAPPED TO THE PROPOSED PLATFORM

Reference	Research focus	Relevance to project	Considerations
Jabeen et al. [1]	Multimodal deep learning	Conceptual basis for handling several modalities	Integration does not guarantee consistent outputs
Kheddar et al. [2]	Automatic speech recognition	Audio-to-text	Noise, domain variation, model resources
Text-to-image literature [3]	Generative image synthesis	Text-to-image	Prompt adherence, quality, latency
Barakat et al. [4]	Expressive speech synthesis	Text-to-audio	Intelligibility, naturalness, language support
Vaishnavi & Narmatha [5]	Video captioning	Video-to-text	Temporal understanding and factual accuracy
Che et al. [6]	Multimodal federated learning	Privacy-related future scope	Not specified as a current platform feature

V. PROPOSED PLATFORM WORKFLOW

The project presentation describes a six-step workflow: (1) the user opens the platform; (2) selects a conversion type; (3) uploads a file or enters text; (4) the system validates and preprocesses the input; (5) the appropriate AI/ML technology is selected; and (6) the output is generated and delivered. The presentation groups the software into image processing, text processing, audio processing, video processing, input processing, and output modules.

A practical implementation can route each request to a dedicated processing path. Image-to-text can use OCR; image-to-audio can combine OCR with TTS; text-to-audio uses speech synthesis; audio-to-text uses ASR; video-to-text can use video captioning or speech transcription depending on the requested output; and video-to-audio may extract the audio track from a video file. The project presentation lists video-to-audio as a conversion service but does not specify whether it means extracting existing audio or generating new audio. That behavior should be defined clearly in the implementation.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 2, October 2026
Impact Factor: 8.2

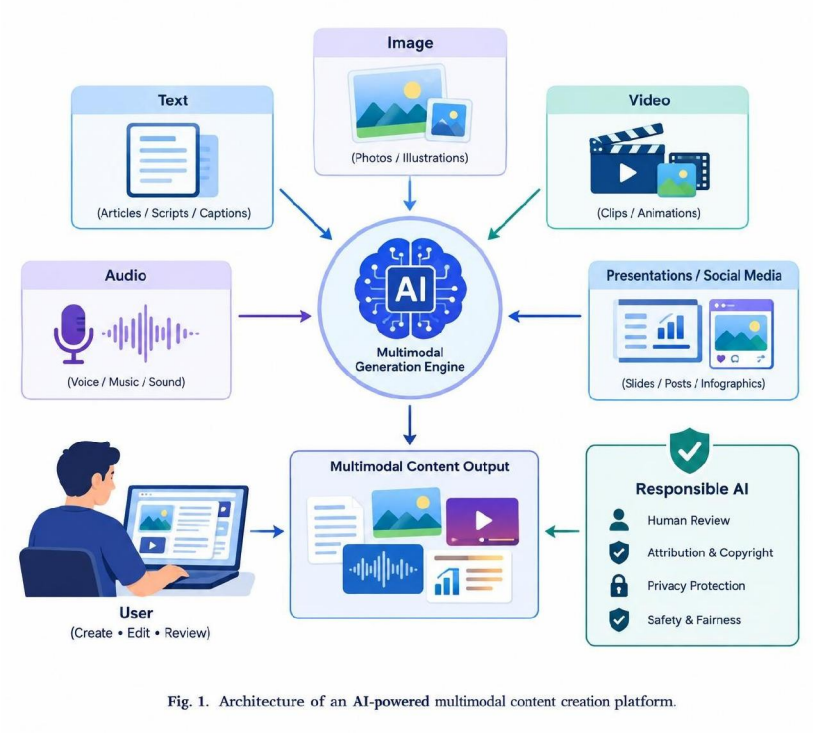


Fig. 1. Architecture of an AI-powered multimodal content creation platform.

Fig. 1 Architecture of an AI-powered multimodal content creation platform.

VI. ADVANTAGES, LIMITATIONS, AND RESEARCH GAPS

The expected advantages identified in the project presentation include a single platform for multiple conversions, less switching between applications, automated processing, a simple interface, and improved accessibility and content reuse. The listed applications include education, content creation, accessibility, general multimedia use, and digital documentation. These are expected benefits; the presentation does not report a user study or measured time savings.

The presentation also identifies current limitations: multilingual support and real-time processing are not yet available, the system is web-only, AI/API services require internet access, and output quality depends on model accuracy. These constraints suggest several evaluation priorities. Each conversion should be tested with representative files and assessed using task-appropriate measures, such as OCR character error, ASR word error, TTS intelligibility, text-to-image prompt adherence, and video-caption relevance. A usability test could also measure task completion time and user satisfaction. Results should be reported only after the platform has been implemented and tested.

Important research gaps include reliable conversion across languages and input conditions, consistent error handling, privacy and file-retention policy, accessibility of the interface, and transparent reporting of model limitations. Future development listed in the presentation includes multilingual support, real-time processing, a mobile application, cloud integration, and more advanced AI models.

VII. CONCLUSION

The reviewed literature provides a research basis for the seven conversion operations proposed in the AI-Powered Multimodal Content Transformation Platform. Multimodal deep learning supplies the overall conceptual framework [1], ASR supports audio-to-text [2], text-to-image generation supports image creation from prompts [3], expressive speech synthesis supports text-to-audio [4], and video-captioning research informs video-to-text [5]. Privacy-aware learning literature offers useful direction for future privacy-sensitive development [6]. The main contribution of the

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 2, October 2026
Impact Factor: 8.2

proposed project is the integration of multiple conversion services into a single, user-friendly web workflow. Further work should implement and test each module, define ambiguous conversion behavior, evaluate output quality and usability, and address multilingual support, latency, privacy, and reliability.

ACKNOWLEDGMENT

The author acknowledges the researchers whose publications inform this review and the project team and guide identified in the supplied project presentation.

REFERENCES

[1]. S. Jabeen, X. Li, M. S. Amin, O. Bourahla, S. Li, and A. Jabbar, "A Review on Methods and Applications in Multimodal Deep Learning," ACM Transactions on Multimedia Computing, Communications, and Applications, vol. 19, no. 2s, Article 76, 2023, doi: 10.1145/3545572.

[2]. H. Kheddar, M. Hemis, and Y. Himeur, "Automatic Speech Recognition Using Advanced Deep Learning Approaches: A Survey," Information Fusion, vol. 109, Article 102422, 2024, doi: 10.1016/j.inffus.2024.102422.

[3]. "Text-to-Image Synthesis With Generative Models: Methods & Challenges," IEEE Access, 2024. The supplied presentation does not provide author names or a DOI for this item; full bibliographic details should be confirmed before final submission.

[4]. H. Barakat, O. Turk, and C. Demiroglu, "Deep Learning-Based Expressive Speech Synthesis: A Systematic Review of Approaches, Challenges, and Resources," EURASIP Journal on Audio, Speech, and Music Processing, vol. 2024, Article 11, 2024, doi: 10.1186/s13636-024-00329-7.

[5]. J. Vaishnavi and V. Narmatha, "Video Captioning—A Survey," Multimedia Tools and Applications, vol. 84, pp. 947–978, 2025 (published online 2024), doi: 10.1007/s11042-024-18886-6.

[6]. L. Che, J. Wang, Y. Zhou, and F. Ma, "Multimodal Federated Learning: A Survey," Sensors, vol. 23, no. 15, Article 6986, 2023, doi: 10.3390/s23156986.