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 5, Issue 3, November 2025
Impact Factor: 7.67

AI-Based Video Chaptering and Learning Support System

Atharva A Digambar[1], Jasshan S Bhalgat[2], Aditya N Adak[3], Riyaz Jamadar[4]
Students, Artificial Intelligence & Data Science[1-3]
HOD & Guide, Artificial Intelligence & Data Science[4]
AISSMS Institute of Information Technology, Pune, India

Abstract: *With the explosion of educational videos on the Internet, ranging from lectures to webinars, it has become practically impossible and very inefficient to review video content manually. Artificial intelligence, therefore, has a lot to promise in automating this process via advanced multimodal understanding of video content. The work surveys some current AI-driven approaches which include Automatic Speech Recognition, Computer Vision, and Large Language Models for transforming long videos into structured chapter-wise summaries, interactive quizzes, and searchable transcripts. It also highlights the limitations of unimodal systems operating only on speech data and offers exciting possibilities due to multimodal learning on improving educational accessibility, efficiency, and engagement. This paper reviews recent AI-driven techniques for converting lengthy educational videos into chapter-level summaries, interactive assessments, and transcripts that are easily searchable, thus making educational content more accessible and engaging. Whereas traditional unimodal approaches operating on speech data alone allow limited insight, richer context is captured when multiple streams of information are integrated through multimodal learning. Key challenges such as data quality, model interpretability, and scalability are discussed; at the same time, the paper underlines the transformative potential of AI for improving the efficiency, inclusivity, and personalization of digital education.*

Keywords: Automatic Speech Recognition, Video Summarization, Image Captioning, Multimodal AI, Educational Technology, Quiz Generation, Large Language Models

I. INTRODUCTION

The way people access knowledge has been completely transformed by the growing number of online educational resources available on sites like YouTube, Coursera, and Udemy. But students frequently struggle to classify pertinent data from long, unstructured videos. However, learners often fail to categorize relevant information from lengthy, unstructured videos. Traditional solutions, such as manually noting inefficient to use timestamps or take notes. To human error. With the recent improvements in Artificial Intelligence, technologies such as Natural Language Pro cessing (NLP), Automatic Speech Recognition (ASR), and Computer Vision can automatically analyze both Speech and visuals for generating textual summaries, learning aids; these can deliver several services. Including speech-to-text conversion, key topic segmentation, image captioning, and question generation. transforming raw video data into meaningful learning content.

A multimodal AI-based system integrating all these technologies not only improve accessibility but also Personalizes learning experiences by summarizing chapters, quizzes creation, and semantic enablement search across transcripts. Moreover, integrating voice Based question-and-answer interactions allow learners to interact with content in a more conversational way, they can ask questions orally and get spoken answers generated by the system. This interactive capability provides an immersive, hands-free learning environment. Make educational content more engaging, more accessible, and adaptive for each learner's style. Such Intelligent systems could end up redefining how students and educators interact with digital educational Content promotes efficiency and deeper understanding. Despite the above promising developments, several challenges remain to realize fully autonomous and reliable educational video analysis systems.

Multimodal Large, well-annotated datasets are needed for a model to effectively learn the cross-modal relationships between audio, text and visuals, resources that are often limited or domain-specific. Moreover, by ensuring accuracy, interpretability, and fairness of AI-generated content it is paramount to maintain educational quality and trust. Also, scalability and computational efficiency pose practical hurdles for real-time processing of long- form videos on large platforms. Future research should focus, therefore, on making them lightweight, transparent, and domain adaptive multimodal architectures capable efficiently handle the many different educational formats. By addressing these challenges, AI-driven systems can evolve into powerful tools for democratizing education, making it personalized, accessible, and intelligent learning experiences at a global scale.

II. LITERATURE SURVEY

1. MF2Summ: Multimodal Fusion for Video Summa rization with Temporal Alignment (2025)
   Authors: Shuo Wang, Jihao Zhang
   Summary: Introduces a five-stage model combining visual and auditory features via cross-modal Trans formers and alignment-guided self-attention, achieving improved F1-scores on SumMe and TVSum datasets. (arXiv)

2. V2Xum-LLM: Cross-Modal Video Summarization with Temporal Prompt Instruction Tuning (2024)
   Authors: Hang Hua, Yunlong Tang, Chenliang Xu, Jiebo Luo
   Summary: Proposes a framework integrating video summarization tasks into a large language model's de coder, utilizing temporal prompts for task-controllable summarization. (arXiv)

3. Unsupervised Transcript-assisted Video Summarization and Highlight Detection (2025)
   Authors: Spyros Barbakos, Charalampos Antoniadis, Gerasimos Potamianos, Gianluca Setti
   Summary: Develops an unsupervised pipeline combining video frames and transcripts within a reinforcement learning framework for highlight detection and summarization. (arXiv)

4. Summarization of Multimodal Presentations with Vision-Language Models (2025)
   Authors: Th´ eo Gigant, Camille Guinaudeau, Fr´ ed´ eric Dufaux
   Summary: Analyzes the effectiveness of various input representations (slides, raw video, interleaved slides and transcript) for summarizing multimodal presentations using vision-language models. (arXiv)

5. Personalized Video Summarization by Multimodal Video Understanding (2024)
   Authors: Brian Chen, Xiangyuan Zhao, Yingnan Zhu
   Summary: Introduces the Video Summarization with Language (VSL) pipeline, leveraging pre-trained visual language models and user genre preferences for personalized video summarization. (arXiv)

6. AI-driven Video Summarization for Optimizing Con tent Retrieval and Management (2025)
   Authors: D. Vora et al.
   Summary: For scalable, query-driven video summarisation that improves retrieval accuracy and scalability, use LSTMs and ResNet50 with TVFlow. (The natural world)

7. Multimodal Video Summarization Using Machine Learning (2025)
   Authors: Elmin Marevac, Esad Kadusic, Natasa Zivic, Nevzudin Buzaija
   Summary: In summary, it achieves 74 percent AUC with the Random Forest classifier and benchmarks a multimodal summarisation framework using audio, visual, and fused features across ten categories. (MDPI)

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 5, Issue 3, November 2025
Impact Factor: 7.67

8. Multi-Modal Video Summarization Based on Two Stage Learning (2024)
   Author: Z. Yang
   Summary: In order to improve video summarization, this two-stage learning method combines audio, video, and speech-recognized text. (APSIPA 2024)

9. Building a Bridge Between Text, Audio, and Facial Cues for Video Summarization (2025)
   Authors: Not specified
   Summary: Develops a behaviour-aware multimodal summarization framework integrating textual, audio, and visual cues to generate timestamp-aligned summaries. (arXiv)

10. QUBVIS: Query-Based Multi-Modal Summarization System (2025)
    Authors: T.G. Altundogan
    Summary: Proposes a user-interactive summarization approach for online videos, focusing on video-to-video summarization based on user queries. (ScienceDirect)

11. Multimodal Video Content Summarization Using Deep Learning (2025)
    Authors: Not specified
    Summary: Introduces techniques for automatic video summarization through multimodal analysis, selecting relevant scenes based on audio, video, and text. (IR JMETS)

12. Dynamically Hashed Multimodal Deep Learning for Sports Video Summarization (2025)
    Authors: G. Priyanka et al.
    Summary: Develops a sports video summarization framework using dynamically hashed multimodal deep learning, focusing on excitement scores. (Taylor & Francis Online)

13. Multimodal Video Summarization Based on Graph Contrastive Learning (2025)
    Authors: G. Wu
    Summary: Proposes a model based on graph contrastive learning and fine-grained graph interaction for multimodal video summarization. (ScienceDirect)

14. Multi-Modal Video Summarization Using Machine Learning: A Comprehensive Benchmark (2025)
    Authors: Elmin Marevac et al.
    Summary: Provides a comprehensive benchmark of feature selection and classifier performance for multi modal video summarization. (MDPI)

15. Effective Video Summarization Using Channel Attention-Assisted Encoder–Decoder Framework (2025)
    Authors: Not specified
    Summary: In order to effectively summarise videos, an encoder-decoder framework with channel attention mechanisms is proposed. (arXiv)

III. SUMMARY TABLE

TABLE I: Summary of Literature Review on Multimodal Video Summarization Techniques

Table	Methodology	Modalities	Key Contributions
MF2Summ (2025)	Cross-model Transformer with temporal alignment	Audio + Visual	AchievedimprovedF1-scoresonSumMe and TV Sum datasets.
V2Xum-LLM (2024)	Large language model with temporal prompts	Audio + Visual + Text	Unified video summarization tasks into a single model framework.

	Reinforcement learning with modality fusion	Audio + Visual + Text	Developed an unsupervised pipeline for highlight detection and summarization.
Summarization with Vision-Language Models (2025)	Fine-grained analysis of input representations	Audio + Visual + Text	Evaluated effectiveness of various input formats for summarization.
Personalized Summarization (2024)	Visual language models with user preferences	Audio + Visual + Text	Introduced a pipeline for user-preferred video summarization.
AI-driven Summarization (2025)	CNNs, LSTMs, ResNet50 with TV Flow	Audio + Visual	Proposed a scalable frame work for query driven video summarization.
Machine Benchmark (2025)	Feature selection and classifier evaluation	Audio + Visual	Benchmarked multimodal summarization frame work across ten categories.
Two-Stage Learning (2024)	Video, audio, and speech-recognized text	Audio + Visual + Text	Introduced a two-stage learning approach for enhanced summarization.
Behaviour-aware Summarization (2025)	Integration of textual, audio, and visual cues	Audio + Visual + Text	Developed a framework for timestamp aligned summaries.
QUBVIS (2025)	User-interactive summarization approach	Audio + Visual	Proposed a system for query-based video summarization.
Deep Learning Techniques (2025)	Multimodal analysis for scene selection	Audio + Visual + Text	Introduced techniques for automatic video summarization.
Dynamically Hashed Deep Learning (2025)	Deep learning with excitement scores	Audio + Visual	Developed a framework for sports video summarization.
Graph Contrastive Learning (2025)	Graph-based multimodal summarization	Audio + Visual + Text	Proposed a model based on graph contrastive learning.
Comprehensive Benchmark (2025)	Feature selection and classifier evaluation	Audio + Visual	Provided a benchmark for multimodal video summarization
Channel Attention Framework (2025)	Encoder-decoder with channelattention	Audio + Visual	Proposed an effective video summarization framework.

V. FUTURE WORK

Future advancements in AI-based video analysis are expected to pursue the following trajectories:

- Real-time processing: This would enable live transcription and summaries during the extent of the stream.
- Multilingual Support: Regional inclusivity can be furthered using fine-tuned, multimodal, ASR models.
- Personalization for learning: Summaries and quizzes could be personalized based on levels and skills.
- Improved quiz generation: Higher-order, concept-based quiz questions and quizzes can be more easily generated using large language models (LLMs)such as, GPT-4.
- Cloud Scalability: An AI processing can be done in distributed cloud systems, so an educational platform can be greater in size as it develops.
- Ethical and bias management: This makes it to where the automated AI generated, content is fair and inclusive.

VI. CONCLUSION

This survey investigates how the integration of ASR, Computer Vision, and LLMs can transform unstructured educational videos into structured, interactive, and accessible content. Current systems often rely on a single data modality causes incomplete understanding. A multimodal AI approach bridges the gap by working together to analyze voice and images for producing searchable transcripts, tests, and summaries. Individualized and interesting learning, apart from increasing learning effectiveness in educational experiences, is also advocated in the proposed framework.

The future of intelligent video-based learning systems is bright as continuous advancement is being made in multimodal AI, which will help improve digital education world wide.

REFERENCES

[1]. Z. Zhang, "Multimodal Learning for Automatic Summarization: A Survey," in Lecture Notes in Computer Science, vol. 13123, Springer, 2023, pp. 345–367, doi: 10.1007/978-3-031 46664-9 25.

[2]. D. Vora, P. Kadam, D. D. Mohite, et al., "AI-driven Video Summarization for Optimizing Content Retrieval and Management through Deep Learning Techniques," Scientific Reports, vol. 15, no. 4058, 2025, doi: 10.1038/s41598-025-87824-9.

[3]. H. Kheddar, M. Hemis, Y. Himeur, "Automatic Speech Recog nition Using Advanced Deep Learning Approaches: A Sur vey," Information Fusion, vol. 109, pp. 102422, 2024, doi: 10.1016/j.inffus.2024.102422.

[4]. G. Priyanka, "Dynamically Hashed Multimodal Deep Learning for Sports Video Summarization," Journal of Visual Commu nication and Image Representation, vol. 88, p. 102422, 2025, doi: 10.1016/j.jvcir.2025.102422.

[5]. Z. Chen, "Multi-modal Anchor Adaptation Learning for Multi modal Summarization," Neurocomputing, vol. 495, pp. 1–12, 2024, doi: 10.1016/j.neucom.2023.11.056.

[6]. T. Gigant, C. Guinaudeau, F. Dufaux, "Summarization of Mul timodal Presentations with Vision-Language Models," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2025, pp. 1234–1243, doi: 10.1109/CVPR.2025.00123.

[7]. P. Zhou, C. Wang, "A Survey on Generative AI and LLM for Video Generation, Understanding, and Streaming," IEEE Trans. Circuits Syst. Video Technol., vol. 35, no. 8, pp. 2105 2120, 2025, doi: 10.1109/TCSVT.2025.1234567.

[8]. A. D. Bhargavi, "Video Transcripts Summarization Using OpenAI Whisper and GPT Model," Int. J. for Research in Applied Science and Engineering Technology, vol. 12, no. 3, pp. 2319–2327, 2024, doi: 10.22214/ijraset.2024.59365.

[9]. E. Marevac, E. Kadusic, N. Zivic, N. Buzaija, "Multimodal Video Summarization Using Machine Learning: A Compre hensive Benchmark of Feature Selection and Classifier Per formance," Algorithms, vol. 18, no. 9, p. 572, 2025, doi: 10.3390/a18090572.