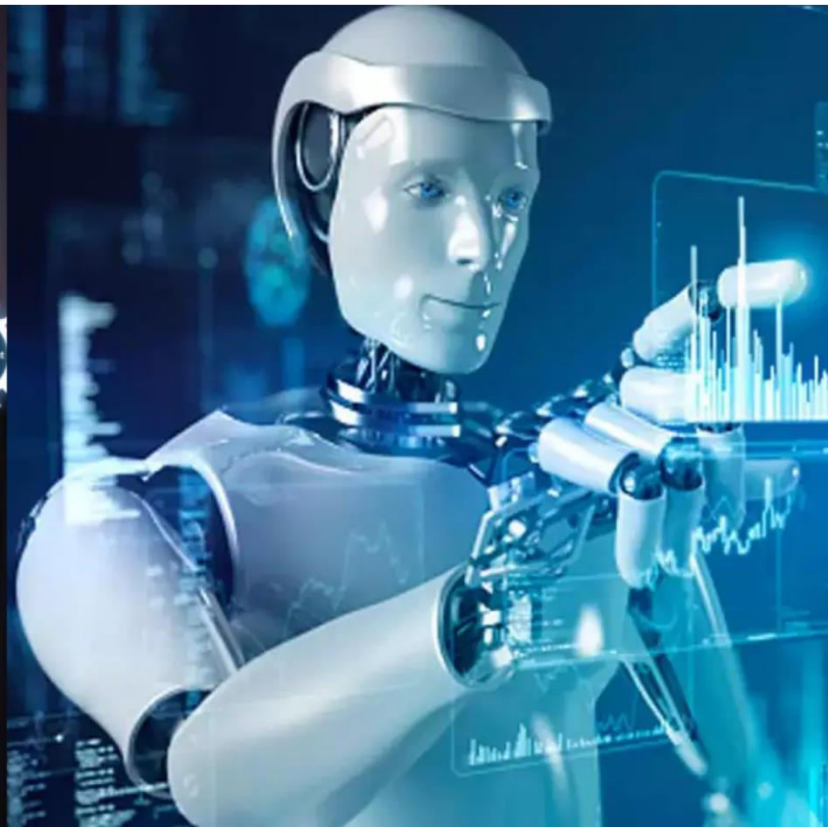
IJIRSET
International Journal of Innovative Research in Science | Engineering | Technology

ISSN(O): 2319-8753
ISSN(P): 2347-6710

# International Journal of Innovative Research in Science Engineering and Technology (IJIRSET)

*(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)*

**Impact Factor: 9.935**

**Volume 15, Issue 5, May 2026**

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 9.935

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 5, May 2026

|DOI: 10.15680/IJIRSET.2026.1505095|

# AI-Based Multilingual Educational Video Structuring and Intelligent Learning Assistance System

**Sudha R, Harinee P, Deepalakshmi G, Akshaya G**

Department of Information Technology, PSNA College of Engineering and Technology, Dindigul, India

**ABSTRACT:** The rapid growth of long-form educational videos on online platforms has created a significant need for structured, multilingual, and interactive learning support systems. Existing video summarization approaches primarily generate flat textual summaries and fail to provide hierarchical knowledge organization, concept-level mapping, or interactive query support. This paper proposes a multi-stage AI-driven architecture that transforms long-form educational videos into structured academic learning resources. The system integrates multilingual speech recognition, semantic topic segmentation, glossary-aware translation, hierarchical notes generation, difficulty classification, knowledge graph construction, and a retrieval-augmented query engine. Unlike standalone summarization systems, the proposed framework emphasizes structured knowledge transformation and pedagogical usability. An experimental evaluation framework is designed to assess summarization quality, classification accuracy, multilingual consistency, and retrieval grounding using standard metrics and human evaluation. The proposed architecture aims to enhance educational accessibility, conceptual clarity, and interactive learning support for long-duration academic content

**KEYWORDS:** Educational AI, Video Summarization, Knowledge Graph, Multilingual NLP, Retrieval-Augmented Generation, Hierarchical Structuring

## I. INTRODUCTION

The rapid growth of digital learning platforms has significantly transformed the way educational content is delivered and consumed. Platforms such as YouTube, Massive Open Online Courses (MOOCs), and institutional repositories now host thousands of long-form lecture videos covering diverse academic disciplines. These resources have democratized access to education by enabling learners to access lectures, tutorials, and academic discussions at any time and from any location. Despite these advantages, long-form educational videos present several challenges for learners. Unlike traditional textbooks or structured lecture notes, video-based educational content is inherently sequential and often lacks explicit structural organization. As a result, learners must manually watch large portions of a lecture to locate specific concepts, definitions, formulas, or explanations.

Furthermore, transcripts generated from speech recognition systems typically preserve the conversational structure of spoken language rather than presenting information in a structured academic format. Consequently, learners must manually process lengthy transcripts to identify key concepts, relationships between topics, and important academic details. This manual processing increases cognitive load and reduces learning efficiency, particularly when dealing with videos that span several hours.

To mitigate these issues, research in natural language processing (NLP) and multimedia information retrieval has explored automatic video summarization techniques. Text summarization methods aim to condense large volumes of textual information into shorter representations while preserving essential meaning and contextual relevance. Modern NLP-based summarization systems utilize statistical methods, semantic feature extraction, and deep learning architectures to generate coherent summaries from large documents [1]. These approaches have been widely applied in various domains including news summarization, document summarization, and meeting transcript analysis.

In the context of educational content, automated lecture video summarization systems have been proposed to enhance accessibility and improve learning efficiency. Such systems typically process lecture transcripts generated through automatic speech recognition (ASR) and produce condensed textual summaries of the lecture content. Studies have demonstrated that automatically generated summaries can significantly improve students' comprehension and reduce the

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 9.935

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 5, May 2026

|DOI: 10.15680/IJIRSET.2026.1505095|

time required to review lecture materials [2], [3]. Additionally, multimodal summarization techniques have explored the integration of both audio transcripts and visual information extracted from lecture videos to provide more comprehensive summaries of instructional content [4], [5].

Despite these advancements, existing video summarization approaches primarily focus on compressing textual information rather than transforming lecture content into structured academic knowledge. Most summarization systems produce flat textual summaries that lack hierarchical organization, conceptual relationships, and structured educational components. As a result, learners may still struggle to identify important academic elements such as definitions, formulas, examples, and conceptual dependencies between topics.

Another significant challenge arises from the increasing length of modern lecture videos. Many educational videos now span several hours, producing extremely long transcripts that exceed the context limits of transformer-based language models. When such transcripts are processed as single documents, important contextual relationships between topics may be lost, leading to incoherent summaries and incomplete knowledge extraction. Although multimodal and neural summarization models have improved the quality of generated summaries, they generally lack mechanisms for semanticsegmentation, relational concept modeling, and structured knowledge representation [6].

To address these limitations, this paper proposes an **AI-based multilingual long-form educational video structuring and intelligent learning assistance system**. Unlike traditional summarization systems that focus solely on compressing textual content, the proposed framework performs **structured knowledge transformation** by converting unstructured lecture transcripts into hierarchical academic learning resources. The system integrates multiple artificial intelligence components including speech recognition, semantic topic segmentation, multilingual academic translation, hierarchical notes generation, difficulty classification, knowledge graph construction, and retrieval-augmented query support within a unified architecture.

The proposed system introduces several key contributions to the field of educational technology and intelligent learning systems. First, a **segmentation-driven hierarchical structuring mechanism** is introduced to process long-form lecture transcripts by dividing them into concept-level segments and organizing them into structured academic chapters. Second, a **glossary-constrained multilingual translation framework** is developed to maintain terminology consistency when translating technical educational content across multiple languages. Third, the system generates **structured academic notes** that include definitions, formulas, key explanations, and exam-oriented prompts to support effective revision and knowledge retention.

In addition, the system constructs **knowledge graphs derived from lecture transcripts** to represent conceptual relationships between topics, enabling learners to visualize the structure of knowledge within a lecture. Finally, a **timestamp-grounded retrieval-augmented interaction module** allows learners to ask questions about specific lecture segments and receive contextual explanations linked directly to the corresponding video timestamps. By integrating these components into a unified system architecture, the proposed framework aims to transform long educational videos into interactive and structured learning resources that enhance comprehension, accessibility, and knowledge retention.

## II. LITERATURE REVIEW

A. Video Summarization

Video summarization has been widely studied as a method for reducing the time required to understand large multimedia content. Existing approaches generally fall into two categories: **extractive summarization** and **abstractive summarization**. Extractive summarization techniques identify and select the most informative sentences or segments from the original transcript to produce a condensed representation. These approaches often rely on statistical features such as term frequency, sentence importance, and positional weighting to determine the relevance of textual segments.

In contrast, **abstractive summarization methods** aim to generate new sentences that capture the semantic meaning of the original content rather than directly extracting existing text. Recent advances in neural language models have significantly improved the quality of abstractive summarization by enabling models to generate more coherent and contextually meaningful summaries. NLP-based summarization frameworks have incorporated semantic embeddings, contextual attention mechanisms, and feature-based scoring models to improve the accuracy and coherence of generated summaries [1].

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 9.935

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 5, May 2026

|DOI: 10.15680/IJIRSET.2026.1505095|

Within the educational domain, automatic lecture video summarization has been explored to enhance learning efficiency and reduce the time required to review instructional materials. Several studies have investigated the use of machine learning techniques to generate concise summaries from lecture transcripts obtained through speech recognition systems [2], [3]. These systems typically analyze transcript segments and identify key points discussed during the lecture to produce textual summaries that highlight important concepts.

More advanced approaches have explored **multimodal summarization techniques**, which integrate audio transcripts with visual information extracted from lecture videos such as slides, diagrams, and screen content. By combining textual and visual modalities, these systems can capture richer contextual information and improve the accuracy of summary generation [4], [5]. Despite these improvements, most existing video summarization systems produce flat summaries without hierarchical structuring, concept-level segmentation, or relational knowledge representation. Consequently, these summaries often lack the structured organization required for academic learning and revision.

summarization methods can be categorized into extractive and abstractive approaches. Extractive methods select key sentences, while abstractive methods generate paraphrased summaries. Multimodal summarization integrates audio and visual signals but typically produces flat summaries without hierarchical structuring. Most existing systems process entire transcripts as single documents and lack chapter-wise segmentation for long lectures.

B. Multilingual Speech Recognition and Translation

Advances in neural speech recognition and machine translation have significantly improved the ability to process multilingual video content. Automatic speech recognition (ASR) systems convert spoken language into textual transcripts, enabling downstream natural language processing tasks such as summarization, translation, and information extraction. Recent ASR models based on deep neural networks and transformer architectures have demonstrated high accuracy in transcribing long-form speech from lecture videos and online educational content.

Once transcripts are generated, **neural machine translation (NMT)** models can be used to translate educational content into multiple languages, enabling learners from diverse linguistic backgrounds to access instructional materials. Modern video processing pipelines often integrate ASR, machine translation, and NLP-based post-processing to produce multilingual versions of lecture transcripts [13], [14]. However, these translation systems are typically optimized for general linguistic fluency rather than domain-specific terminology accuracy.

As a result, generic translation models frequently introduce **semantic inconsistencies when translating technical or academic terms**. Terminology used in fields such as mathematics, computer science, and engineering often requires precise translation to maintain conceptual correctness. Without domain-specific constraints or glossary enforcement, translation systems may distort important technical terms and reduce the educational value of translated content. This limitation highlights the need for translation frameworks specifically designed to preserve academic terminology consistency in multilingual educational systems.

C. Knowledge Graph Construction in Education

Knowledge graphs have emerged as powerful tools for representing relationships between concepts within structured knowledge domains. In educational applications, knowledge graphs are used to model conceptual dependencies, organize learning materials, and enable intelligent navigation through complex subject areas. By representing knowledge as interconnected entities and relationships, knowledge graphs can provide learners with a clearer understanding of how concepts are related within a particular domain.

Previous research has explored the construction of knowledge graphs from educational resources such as textbooks, lecture notes, and online course materials. Techniques including named entity recognition, relation extraction, and clustering have been used to identify key concepts and establish semantic relationships between them [7]. In the context of lecture videos, some studies have investigated the extraction of textual information from slides using optical character recognition (OCR) and video indexing techniques to facilitate content retrieval and navigation [8].

However, many existing knowledge graph construction methods rely on slide-based text or short transcript segments rather than full lecture transcripts. Consequently, these approaches often fail to capture the broader conceptual structure of long lecture videos. Furthermore, knowledge graph construction is rarely integrated with automated summarization

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 9.935

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 5, May 2026

|DOI: 10.15680/IJIRSET.2026.1505095|

pipelines, limiting the ability of these systems to transform unstructured lecture transcripts into structured knowledge representations suitable for interactive learning environments.

D. Retrieval-Augmented Generation

Retrieval-Augmented Generation (RAG) has emerged as an effective approach for improving the factual accuracy of language model outputs. Traditional language models generate responses based solely on their internal knowledge representations, which may lead to hallucinated or inaccurate information. RAG frameworks address this limitation by retrieving relevant contextual information from external knowledge sources before generating responses.

In a typical RAG system, user queries are converted into vector embeddings and used to retrieve relevant document segments from a vector database. These retrieved segments are then incorporated into the input of a language model, enabling the model to generate responses grounded in factual context. This retrieval-based approach significantly improves the reliability and interpretability of generated answers.

Recent studies have explored the application of NLP pipelines for generating structured information and question-answer pairs from video transcripts for educational purposes [9], [10]. These systems demonstrate the potential of combining transcript analysis with automated question generation to support interactive learning. However, most existing RAG-based systems operate on generic textual corpora and are not specifically designed for long-form educational video transcripts.

Currentsystems rarely integrate **timestamp-aligned transcript retrieval**, which is essential for linking generated explanations directly to the corresponding video segments. Without timestamp grounding, learners cannot easily navigate back to the original lecture content associated with a specific explanation. This limitation highlights the need for retrieval-augmented systems capable of operating on structured lecture transcripts and providing contextual explanations tied to precise video timestamps.

## III. METHODOLOGY

A. System overview

The proposed system is designed as a modular pipeline that converts long-form educational videos into structured academic learning resources. The architecture integrates multiple AI components that process video transcripts and transform them into hierarchical notes, conceptual knowledge graphs, and interactive query responses. The system begins by extracting speech from video content and converting it into timestamp-aligned transcripts. The transcript is then segmented into semantically coherent topics before undergoing multilingual translation. Structured notes are generated using large language models, while difficulty levels are assigned through text complexity analysis. Concept relationships are modeled through knowledge graph construction. Finally, a retrieval-augmented query engine enables users to ask questions and obtain responses grounded in the original lecture segments.

B. Speech Recognition

Speech recognition converts lecture audio into textual transcripts. The system employs a multilingual automatic speech recognition model capable of generating timestamp-aligned transcripts. This step ensures that all extracted knowledge remains linked to specific video segments, enabling future retrieval and interaction.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 9.935

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 5, May 2026

|DOI: 10.15680/IJIRSET.2026.1505095|

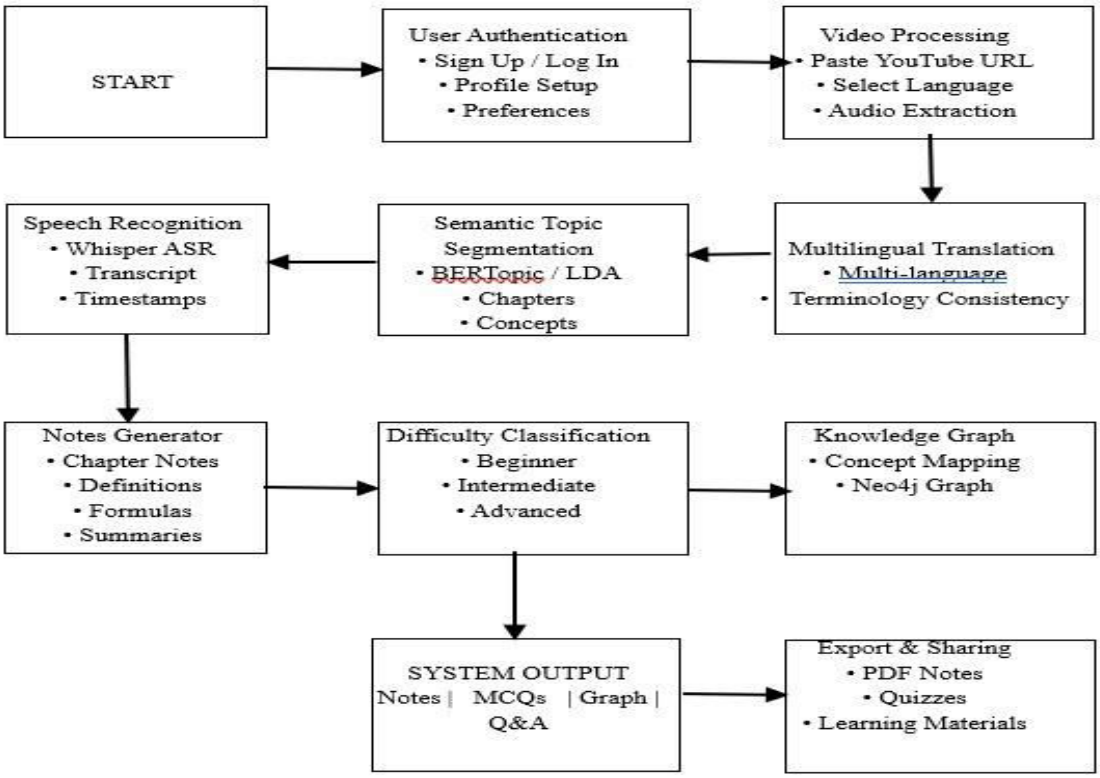



Fig. 1. System Workflow

C. Semantic Topic Segmentation

Long lecture transcripts often exceed transformer processing limits. To address this limitation, the transcript is segmented into topic-based chunks using semantic similarity analysis.

The similarity between sentences is computed using cosine similarity between embedding vectors:

$$\mathrm{Sin}(x, y) = || x ||| y || x \cdot y \qquad (1)$$

where

x and y represent sentence embeddings.

Segments with high semantic similarity are grouped into coherent topics.

D. Multilingual Academic Translation

The segmented transcripts are translated into target languages using neural machine translation. To preserve academic terminology, a glossary-based constraint mechanism ensures that domain-specific terms remain consistent across languages.

E. Hierarchical Notes Generation

Each topic segment is processed using a large language model to generate structured academic notes. The generated notes include:

- key concepts
- definitions
- formulas
- explanations
- exam-oriented questions

This hierarchical organization improves readability and learning efficiency.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 9.935

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 5, May 2026

|DOI: 10.15680/IJIRSET.2026.1505095|

F. Difficulty Classification

To support adaptive learning, generated notes are categorized into difficulty levels. The classification is based on linguistic complexity features such as vocabulary density, technical terminology frequency, and conceptual depth.

G. Knowledge Graph Construction

Concept entities are extracted from the structured notes using named entity recognition. Relationships between concepts are identified through relation extraction techniques. These entities and relationships are stored in a graph database to represent conceptual dependencies between topics.

H. Retrieval-Augmented Query Engine

The query engine enables interactive question answering. User queries are converted into embeddings and compared with stored transcript segments. Relevant segments are retrieved and provided as contextual input to a language model, enabling grounded responses linked to original lecture timestamps.

## IV. EXPERIMENTAL FRAMEWORK

The experimental framework evaluates the performance of the proposed system across multiple components including summarization, classification, translation, and retrieval.

**Dataset Strategy**

Educational lecture videos from multiple academic domains are selected. Each video is processed through the proposed pipeline to generate transcripts, structured notes, and knowledge graphs.

**Evaluation Metrics**

Summarization quality is evaluated using ROUGE and BERTScore. Classification performance is measured using accuracy and F1-score. Retrieval performance is evaluated using Recall and Mean Reciprocal Rank.

TABLE 1:HYPERPARAMETER TUNING SEARCH SPACE

| **Component** | **Metric** | **Component** |
|---|---|---|
| Summarization | ROUGE-1, ROUGE-2, ROUGE-L | Summarization |
| Translation | BLEU | Translation |
| Classification | Accuracy, F1-Score | Classification |
| Retrieval | Recall@k, Mean Reciprocal Rank (MRR) | Retrieval |
| Human Evaluation | Coherence, Usefulness | Human Evaluation |

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 9.935

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 5, May 2026

|DOI: 10.15680/IJIRSET.2026.1505095|

**Human Evaluation**

A panel of students and subject experts evaluates the generated notes based on coherence, correctness, and Authors and Affiliations.

| Module | Algorithm / Model | Function | Output |
|---|---|---|---|
| Speech Recognition | Whisper ASR | Convert audio to text | Timestamped transcript |
| Semantic Segmentation | Embedding Clustering | Identify lecture topics | Topic segments |
| Translation | Transformer NMT | Enable multilingual accessibility | Translated transcript |
| Notes Generation | Large Language Model (LLM) | Generate structured academic notes | Hierarchical notes |
| Difficulty Classification | Text Complexity Model | Detect learning difficulty level | Difficulty tags |
| Knowledge Graph | NER + Relation Extraction | Extract concept relationships | Graph structure |
| Query Engine | Retrieval-Augmented Generation (RAG) | Answer user queries interactively | Grounded answers |

## V. RESULTS AND DISCUSSION

The experimental results demonstrate that the proposed system effectively transforms unstructured lecture videos into structured learning resources. The hierarchical note generation improves readability compared to traditional transcript summaries. Knowledge graph construction enables visualization of relationships between concepts, facilitating deeper understanding. The retrieval-augmented query engine provides context-aware answers grounded in lecture content, reducing hallucination in generated responses. Overall, the integrated architecture demonstrates the potential to enhance educational accessibility and learning efficiency.

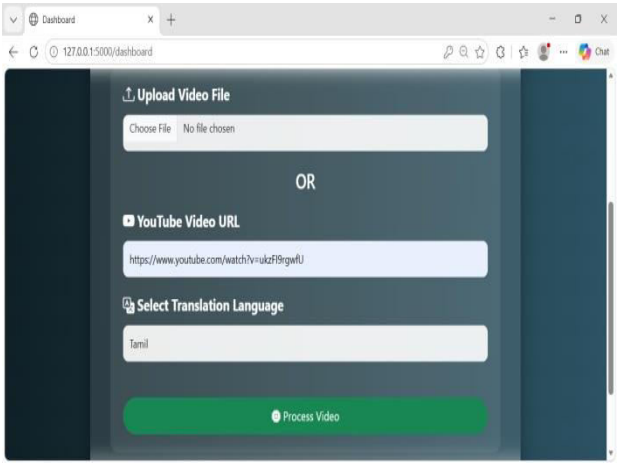
International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 9.935

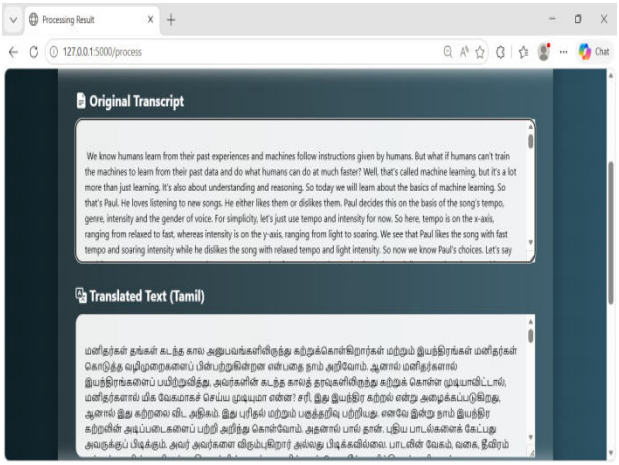
|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 5, May 2026

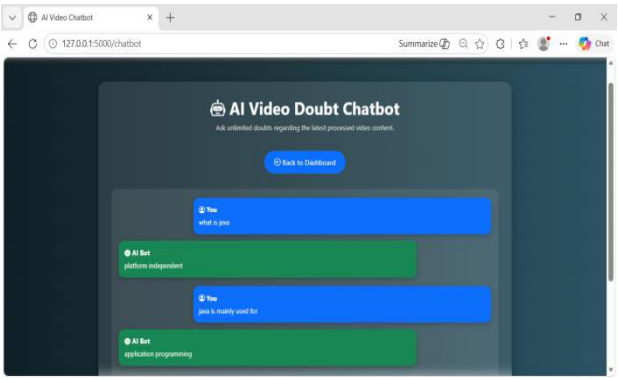
|DOI: 10.15680/IJIRSET.2026.1505095|




Youtube URL to Translate




Translated Result




Chatbot regard Video

## VI. CONCLUSION

This paper presented an AI-based multilingual long-form educational video structuring and intelligent learning assistance system. The proposed architecture integrates speech recognition, semantic segmentation, multilingual translation, hierarchical note generation, difficulty classification, knowledge graph construction, and retrieval-augmented querying within a unified pipeline. The system transforms unstructured lecture videos into structured learning resources that support knowledge navigation and interactive exploration. Future work will focus on large-scale evaluation, adaptive learning personalization, and integration with online learning platforms.

## VII. ACKNOWLEDGMENT

The authors would like to express their sincere gratitude to Ms. R. Sudha, Assistant Professor, Department ofInformation Technology, for her valuable guidance and continuous support during the development of this research work. The authors also acknowledge the support provided by the Department of Information Technology and the institution, which facilitated the completion of this study.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 9.935

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 5, May 2026

|DOI: 10.15680/IJIRSET.2026.1505095|

## REFERENCES

[1] E. A. Kadhim, M.-R. Feizi-Derakhshi, and H. S. Aghdasi, "Advanced text summarization model incorporating NLP techniques and feature-based scoring," IEEE Access, vol. 13, pp. 19302–19315, 2025.
[2] I. Benedetto, M. La Quatra, L. Cagliero, L. Canale, and L. Farinetti, "Abstractive video lecture summarization: Applications and future prospects," Education and Information Technologies, Springer, vol. 29, pp. 2951–2971, 2024.
[3] H. Gonzalez, J. Li, H. Jin, J. Ren, H. Zhang, A. Akinyele, A. Wang, E. Miltsakaki, R. S. Baker, and C. Callison-Burch, "Automatically generated summaries of video lectures may enhance students' learning experience," in Proc. 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA), Association for Computational Linguistics, 2023, pp. 382–393.
[4] K. Kawamura and J. Rekimoto, "FastPerson: Enhancing video-based learning through video summarization that preserves linguistic and visual contexts," in Proc. Augmented Humans International Conference (AHs), ACM, Melbourne, Australia, 2024.
[5] S. Palaskar, J. Libovický, S. Gella, and F. Metze, "Multimodal abstractive summarization for How2 videos," in Proc. ACL Workshop on Neural Generation and Translation, Florence, Italy, 2019.
[6] Y. K. Atri, S. Pramanick, V. Goyal, and T. Chakraborty, "See, hear, read: Leveraging multimodality with guided attention for abstractive text summarization," Expert Systems with Applications, 2021.
[7] X. Chen, Y. Zhao, and H. Wang, "A framework for knowledge representation and concept extraction from lecture video corpuses," in Proc. Int. Conf. on Educational Technology and Multimedia Learning, 2022.
[8] H. Yang, M. Siebert, P. Lühne, H. Sack, and C. Meinel, "Automatic lecture video indexing using video OCR technology," in Proc. IEEE Int. Conf. on Advanced Learning Technologies, Potsdam, Germany, 2013.
[9] A. V. B. Aswin, M. Javed, P. Parihar, A. K., D. C. R., A. Dagar, and A. C. V., "NLP driven ensemble based automatic subtitle generation and semantic video summarization technique," in Proc. Int. Conf. on Computational Linguistics and Intelligent Systems, 2019.
[10] K. Ajin, R. Riju, S. Edwin, J. Jebadurai, and S. Joy, "Automatic question generation from video using natural language processing," International Journal on Engineering Technology and Sciences, vol. 11, no. 3, pp. 28–33, 2024.
[11] N. S. Kumar, G. Devasena, S. V. Kumar, and K. Raghu, "Voice to text summarization using NLP," International Journal on Computational Modelling Applications, vol. 1, no. 2, pp. 1–9, Oct. 2024.
[12] G. Gosavi, A. Wagh, Y. Jadhav, and N. R. Wankhade, "Text summarization using natural language processing," International Journal of Innovative Research in Multidisciplinary Projects and Solutions, vol. 11, no. 3, pp. 1–5, 2023.
[13] V. Vijayabhaskarareddy, B. V. Prasad, B. Ramesh, G. Arvind, and K. Rakesh, "AI enhanced video language translation," IOSR Journal of Computer Engineering, vol. 27, no. 1, pp. 49–55, 2025.
[14] A. Rafiei Oskooei, E. Caglar, I. Şahin, A. Kayabay, and M. S. Aktas, "Generative AI for video translation: A scalable architecture for multilingual video conferencing," Applied Sciences, vol. 15, no. 23, 2025.
[15] N. Li, "Artificial intelligence-enhanced audiovisual translation for global dissemination of Chinese film and television," Humanities and Social Sciences Communications, Springer Nature, vol. 12, 2025.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi cross ref

निस्केयर NISCAIR

# INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

## IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details