International Journal of Electrical, Electronics and Computer Systems
ISSN: 2347-2820
Volume 15 Issue 01 (2026) | Pages: 134-141

Video to Text Summarizer with Highlights and Quiz Generator

Jalindar Nivrutti Ekatpure[1], Samiksha Jagadale[2], Snehal Kargal[3], Sakshi Lavand[4], Sanika Yadav[5]

[1,2,3,4,5]Department of Computer Engineering, S. B. Patil College of Engineering, Indapur, Pune, Maharashtra, India
[1]*j.ekatpure@gmail.com*, [2]*jagadalesamiksha0605@gmail.com*, [3]*snehalkargal0@gmail.com*, [4]*sakshilawand98@gmail.com*, [5]*sanikayadav926@gmail.com*

Peer Review Information

***Type**: Article*
***Received**: 24 March 2026*
***Revised**: 09 April 2026*
***Accepted**: 27 May 2026*
***Published**: 06 June 2026*

Abstract

The increasing use of video-based learning has created a need for efficient content understanding and extraction. This paper presents Video to Text Summarizer with Highlights and Quiz Generator, an AI-powered system that converts video content into structured learning materials such as transcripts, summaries, highlights, quizzes, and timelines. The system uses offline speech recognition (Vosk) and Google Gemini AI to process and analyze video data. It reduces the time required to understand long videos and enhances user engagement through interactive outputs. The proposed solution provides a scalable and efficient approach for improving digital learning and content accessibility.

Keywords: Video Summarization; Speech-to-Text; Natural Language Processing; Artificial Intelligence; Quiz Generation; Educational Technology; Automated Learning.

How to Cite This Article

Ekatpure, J. N., Jagadale, S., Kargal, S., Lavand, S., & Yadav, S. (2026). Video to text summarizer with highlights and quiz generator. *International Journal of Electrical, Electronics and Computer Systems*, 15(1), 134–141.

Introduction

The rapid growth of digital learning platforms has significantly increased the availability of video-based educational content, making it a primary source of knowledge acquisition. However, users often face challenges such as lengthy video durations, difficulty in identifying key concepts, and lack of structured learning materials. Traditional methods of consuming video content require users to watch entire videos, which is time-consuming and may reduce engagement and learning efficiency.

To address these challenges, this project introduces an integrated AI-powered system called Video to Text Summarizer with Highlights and Quiz Generator, which utilizes speech recognition and natural language processing techniques to convert video content into structured learning outputs. The system is designed to automatically generate transcripts, summaries, highlights, quizzes, and timelines, enabling users to quickly understand and revise important concepts without watching the full video.

The proposed system enhances learning productivity by leveraging offline speech-to-text technology and advanced AI models to analyze and process video data. By extracting meaningful insights and presenting them in an interactive format, the platform allows users to focus on essential information, improve retention, and save time. This AI-driven approach ensures that learners receive accurate and structured content, helping them make better use of educational resources. By integrating modern technologies with digital learning, the system bridges the gap between raw video content and intelligent knowledge extraction, promoting efficient, accessible, and engaging learning experiences for a wide range of users.

Literature Survey

1. Video Summarization Techniques: A Comprehensive Review – Toqa Alaa et al. (2024): Explores deep learning-based multimodal approaches; future scope in scalable, real-time summarization.
2. Video-to-Text Summarization using NLP – Prerna Mishra et al. (2023): Uses ASR and SpaCy entity extraction; emphasizes multilingual summarization challenges.
3. Video Transcript Summarization Using BERT – Iswarya M. et al. (2023): Applies BERT for large transcript summarization; reinforcement learning suggested for improvement.
4. Video Transcript Summarizer – Ilampiray P. et al. (2023): Uses YouTube Transcript API + BERT for multilingual summarization with Flask integration.
5. Video Summarization using NLP – Raghav Malu et al. (2023): Combines transcript extraction, transformer models, and ROUGE evaluation for education/surveillance.
6. Video to Text Summarisation and Timestamp Generation – Dhiraj Shah et al. (2022): Uses CNN + GRU with attention for event detection; useful in long video navigation.
7. Automatic Generation of MCQs from Educational Texts in Hindi – Shweta Yadav et al. (2021): Employs BERT embeddings for question generation in Hindi; suggests expansion to other Indian languages.
8. Video Summarization Using Highlight Detection and Deep Learning – Sarvesh Kolhe et al. (2020): Uses CNN + LSTM for highlight detection; scope in real-time summarization.
9. Video Summarization using Deep Semantic Features – Mayu Otani et al. (2016): Employs DNN to capture semantic meaning in videos; suggests improvements in segmentation and retrieval.
10. Other related NLP/EdTech works (from survey list) – Focus on multimodal learning, abstraction accuracy, and e-learning integration.

Limitations of Existing Work

- Existing methods struggle with real-time summarization of long videos.
- Multilingual support is limited in current summarization systems.
- Quiz generation research is underdeveloped, especially for non-English languages.
- Lack of cloud integration and scalability in many approaches.
- Summarization models face hallucination and coherence issues.

Problem Statement

With the increasing volume of video content in education, training, and professional fields, users face challenges in efficiently extracting key information. Manually watching long videos is time-consuming and ineffective. There is a need for an automated system that converts video content into concise text summaries, highlights important parts, and generates quizzes to enhance learning outcomes.

Proposed System

The proposed system, Video to Text Summarizer with Highlights and Quiz Generator, is an advanced AI-powered platform designed to

assist users in efficiently extracting meaningful information from video content by leveraging speech recognition and natural language processing technologies. It aims to address key challenges such as time-consuming video consumption, lack of structured learning material, and limited interactivity in existing systems. By providing automated content extraction, intelligent summarization, and interactive learning features, the system enhances user engagement and improves learning efficiency.

One of the core features of this system is video-to-text transcription, which utilizes offline speech recognition (Vosk) to convert video audio into accurate textual transcripts. This allows users to access video content in text form without relying on internet-based APIs, ensuring better privacy and offline usability. Another important component is AI-based summarization, which uses advanced language models such as Google Gemini AI to generate concise and meaningful summaries from transcripts. This enables users to quickly understand key concepts without watching the entire video, saving time and effort.

The system also includes highlight extraction, where important keywords and key points are identified from the video content. These highlights help users focus on essential information and improve revision efficiency. Additionally, the platform generates interactive quizzes (MCQs) based on the video content, allowing users to test their understanding and reinforce learning. To further enhance usability, the system provides a timeline segmentation feature, which divides the video into meaningful segments. This enables users to navigate directly to specific parts of the video, improving accessibility and user experience.

The platform is built as a full-stack web application using modern technologies such as Next.js, React, and Node.js, with MongoDB for data storage. It includes secure authentication, user dashboards, and history tracking, ensuring a personalized and seamless experience. The user interface is designed to be intuitive and user-friendly, making it accessible even for users with minimal technical knowledge. To maintain system efficiency, the platform incorporates optimized processing techniques, efficient data handling, and scalable architecture to support large volumes of video content. By integrating transcription, summarization, highlight extraction, quiz generation, and timeline navigation into a single unified system, This system provides a comprehensive solution for modern digital learning.

By combining AI-driven insights with interactive features, the proposed system bridges the gap between raw video content and structured knowledge. It promotes efficient learning, improves content accessibility, and provides a scalable solution for students, educators, researchers, and professionals.

System Requirements

Database Requirements

- MongoDB (Mongoose) Database

Software Requirements (Platform Choice)

- Operating System: Windows 10 / Linux
- Coding Language: JavaScript (Node.js), Python
- Frontend: React.js, Next.js, Tailwind CSS
- Backend: Node.js (Next.js API Routes)
- IDE: VS Code
- Web Browser: Google Chrome
- AI Tools: Google Gemini AI
- Speech Recognition: Vosk (Offline)
- Video Processing: FFmpeg, MoviePy, Pydub

Hardware Requirements

- System – Intel i5 / Ryzen 5 or above
- RAM – 8 GB (minimum)
- Hard Disk – 512 GB HDD/SSD
- Keyboard – Standard Windows Keyboard
- Mouse – Two or Three Button Mouse
- Monitor – 15-inch or higher display
- Audio Input – Microphone (for testing speech input)

Methodology

The development of the Video to Text Summarizer with Highlights and Quiz Generator (NeuroStream) follows a structured methodology to ensure efficiency, accuracy, and scalability. The methodology consists of multiple phases, including data acquisition, preprocessing, AI model processing, system integration, and deployment. Each phase contributes to the overall functionality and effectiveness of the system.

Data Collection and Preprocessing

The system processes video data as input and converts it into structured textual information. The data handling process is categorized into two major types:

Video Input Data

- Video files or YouTube URLs are provided by the user as input.
- Videos are downloaded (if URL is provided) using tools like yt-dlp.
- The system extracts audio from the video using FFmpeg.
- Audio is preprocessed (converted to mono, 16kHz format) to improve transcription accuracy.

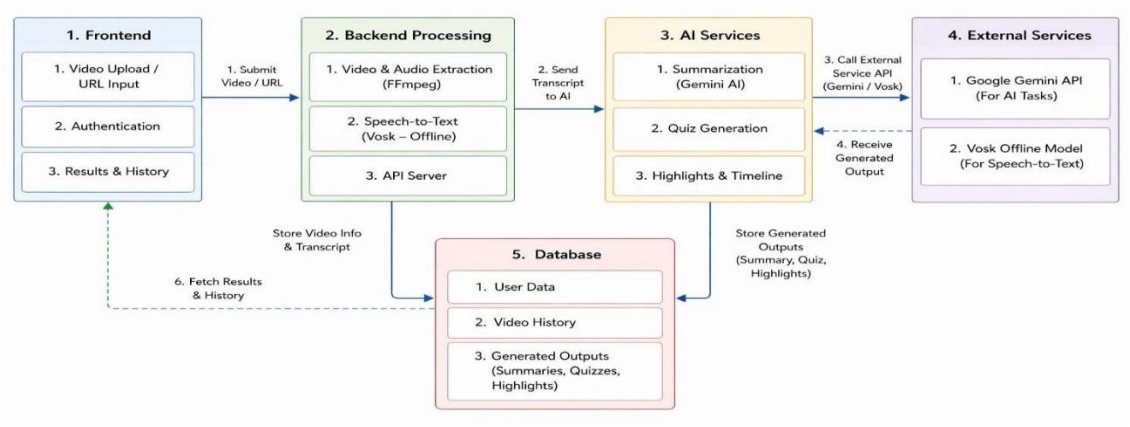

Fig. 1. *System Architecture*

Text Processing Data

- The extracted audio is converted into text using offline speech recognition (Vosk).
- The generated transcript is cleaned to remove noise, filler words, and errors.
- Text normalization techniques are applied to improve readability and processing.

AI Model Processing

The system utilizes AI-based models to generate structured learning outputs:

Speech-to-Text Model

- The Vosk offline speech recognition model is used to convert audio into text.
- It processes audio frames and generates accurate transcripts without internet dependency.
- The output transcript is stored for further processing.

Content Generation Model

Google Gemini AI is used for advanced text processing and analysis.

The model generates:

- Summary (5–7 key points)
- Highlights (important keywords)
- Quiz (multiple-choice questions)
- Timeline (video segmentation)
- NLP techniques ensure context-aware and meaningful output generation.

System Development and Integration

The system is developed as a full-stack web application with the following layers:

User Interface Layer

- A responsive web application is developed using React.js and Next.js.
- Users can upload videos, view transcripts, summaries, and quizzes.
- Dashboard provides history tracking and interactive UI components.

Application Logic Layer

- Backend is implemented using Node.js and Next.js API routes.
- Handles video processing, transcription, and AI requests.

- Manages authentication, user requests, and data flow between modules.

Data Storage Layer

- MongoDB database stores user data, video history, and generated outputs.
- Efficient schema design ensures fast retrieval and scalability.

External Services Integrat

- Google Gemini AI API is integrated for summarization, quiz generation, and content analysis.
- Vosk offline model is used for speech recognition without internet dependency.
- FFmpeg is used for audio extraction and video processing.

System Workflow

User Workflow

1. User logs in and uploads a video file or provides a YouTube URL.
2. The system extracts audio and generates a transcript using Vosk.
3. The transcript is processed by Gemini AI to generate summary, highlights, quiz, and timeline.
4. The results are displayed on the dashboard for user interaction.
5. Users can view, download, and revisit previous results.

Result Discussion

The Video to Text Summarizer with Highlights and Quiz Generator (NeuroStream) system was developed and tested to evaluate its accuracy, efficiency, and overall user experience. The speech-to-text module, implemented using the Vosk offline model, demonstrated reliable performance in generating accurate transcripts from video audio, even without internet dependency. The transcription accuracy was satisfactory for educational and general-purpose videos, although minor errors were observed in cases of background noise or unclear speech.

The AI-based content generation module, powered by Google Gemini AI, produced meaningful and concise summaries, typically consisting of 5–7 key points. The system successfully extracted relevant highlights in the form of keywords and generated multiple-choice questions that effectively reflected the video content. These features significantly improved content understanding and user engagement. The timeline segmentation feature also enabled users to navigate videos efficiently by dividing them into meaningful sections.

The system's performance was efficient, with processing time remaining within acceptable limits for short to medium-length videos. The integration of FFmpeg for audio extraction and optimized backend processing ensured smooth and fast execution of tasks. The React and Next.js-based web interface was found to be user-friendly, allowing users to بسهولة upload videos, view results, and interact with generated outputs. Secure authentication and data storage using MongoDB ensured user data privacy and reliability.

The overall impact of the system was positive, as it reduced the time required to understand video content and provided an interactive learning experience through summaries, quizzes, and highlights. However, certain challenges were identified, such as reduced transcription accuracy for noisy audio, dependency on AI models for content quality, and performance limitations for very long videos. Future improvements can include enhancing speech recognition accuracy, supporting multiple languages, enabling real-time processing, and optimizing system performance for large-scale applications.

Results/ Outputs

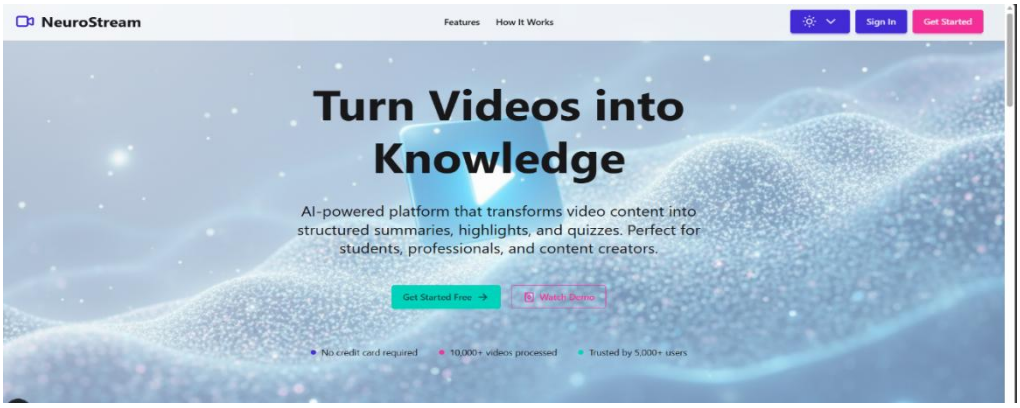

Fig. 2. *NeuroStream Home Page Interface*

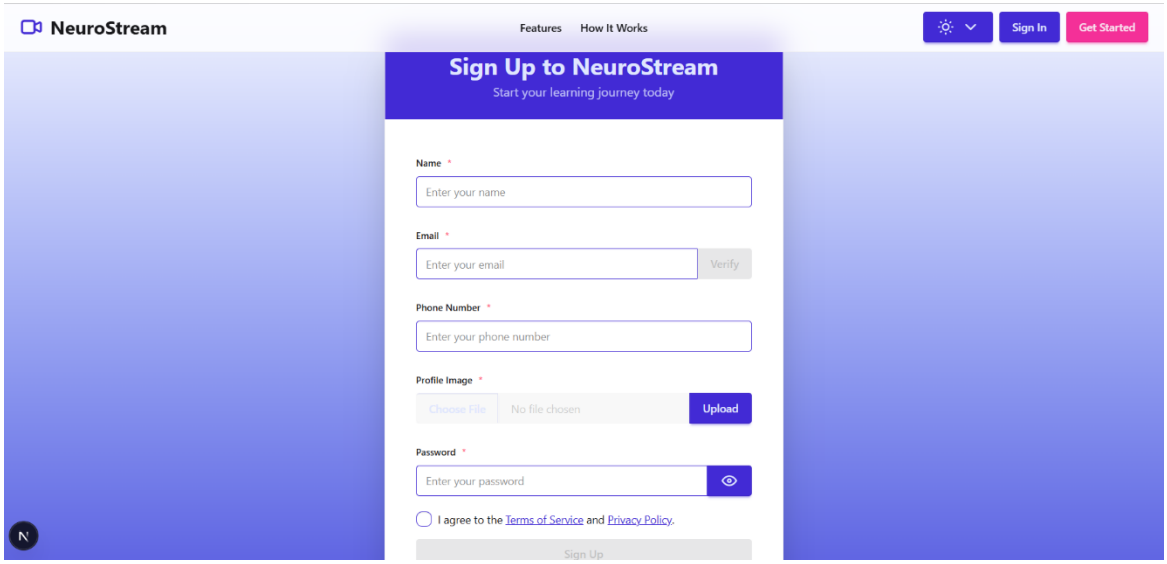

Fig. 3. *Sign up page*

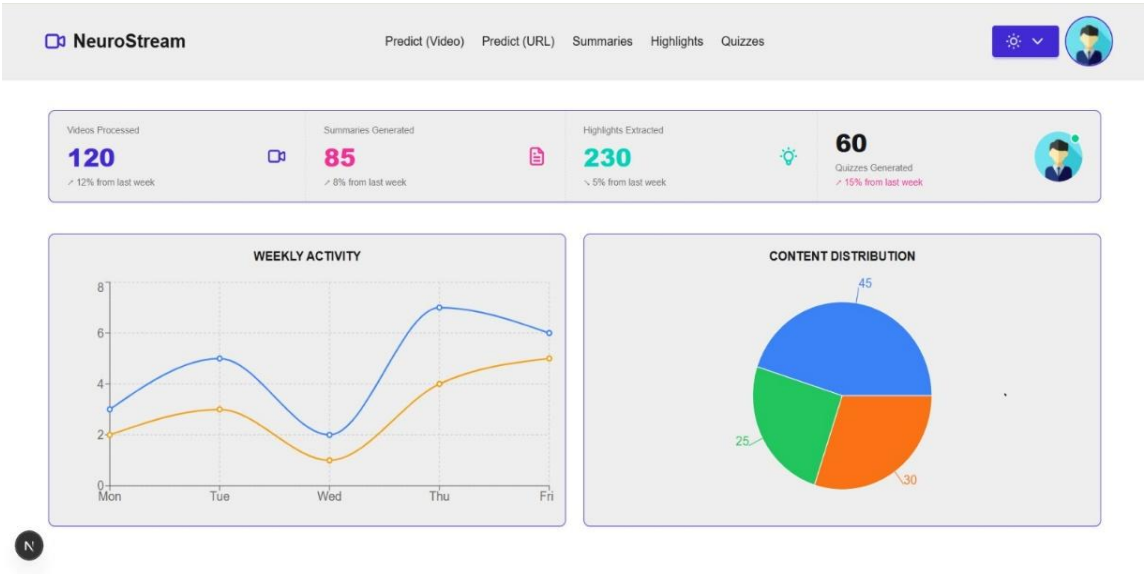

Fig. 4. *Dashboard and Analytics View*

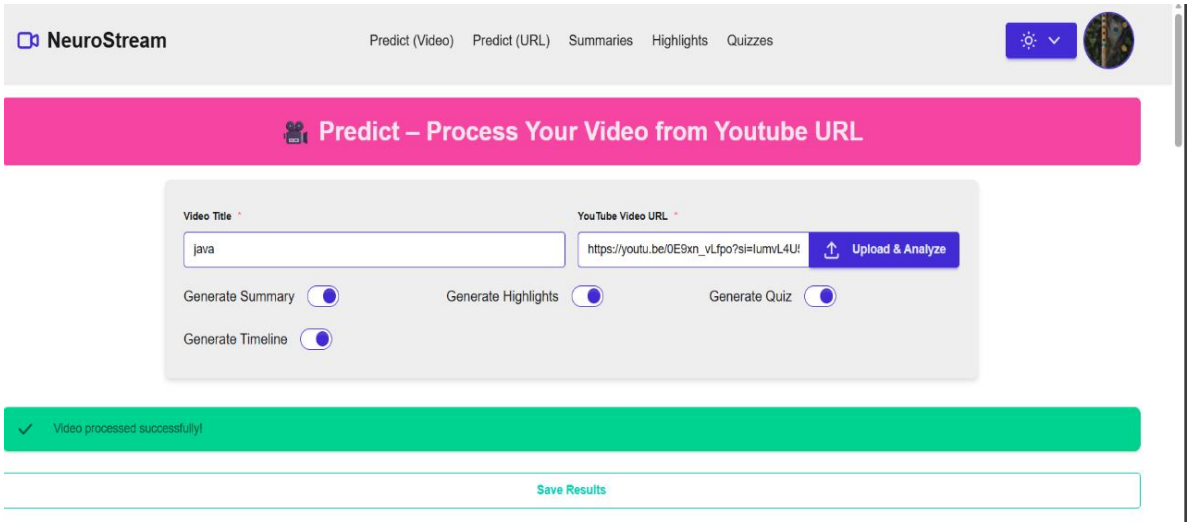

Fig. 5. *Video Processing*

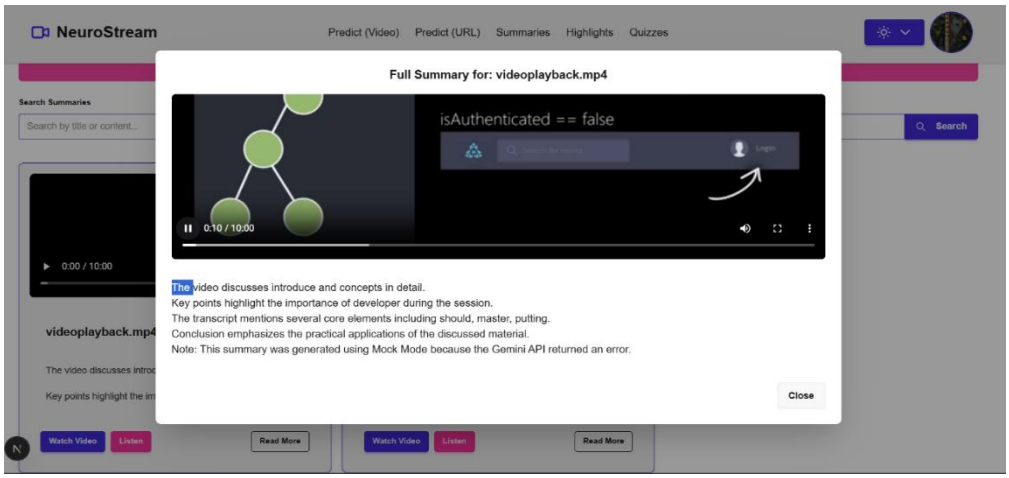

Fig. 6. *Summary*

Fig. 7. *Highlights*

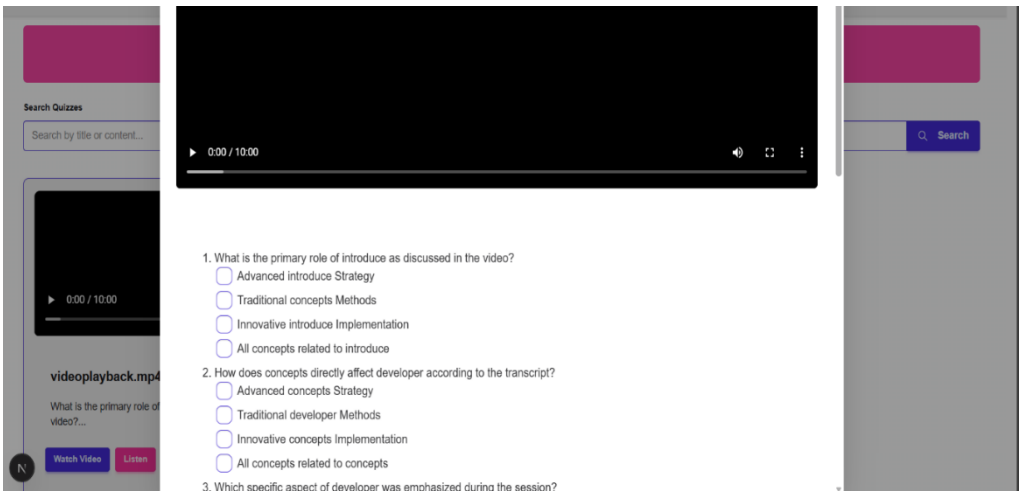

Fig. 8. *Quizzes*

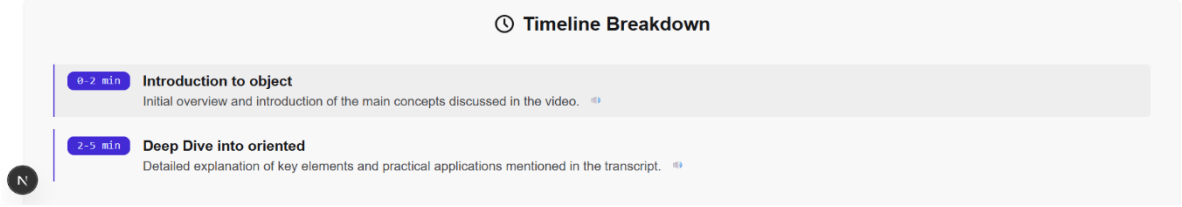

Fig. 9. *Timeline*

Conclusion

The Video to Text Summarizer with Highlights and Quiz Generator (NeuroStream) successfully integrates speech recognition, natural language processing, and modern web technologies to transform video content into structured and interactive learning material. The speech-to-text module effectively converts video audio into accurate transcripts using the Vosk offline model, enabling reliable processing without internet dependency. The AI-powered content generation module, utilizing Google Gemini AI, produces concise summaries, relevant highlights, quizzes, and timeline segmentation, helping users quickly understand key concepts and improve learning efficiency. The web-based platform, developed using React.js and Node.js, provides a user-friendly and interactive interface, allowing users to upload videos, view results, and access previously generated content. Secure authentication and database management ensure data privacy and efficient storage of user activity and outputs. The system significantly reduces the time required to consume video content while enhancing engagement through interactive features such as quizzes and highlights. Despite its effectiveness, certain challenges remain, including handling noisy audio for accurate transcription, improving performance for long-duration videos, and ensuring consistent quality of AI-generated outputs. Future enhancements may include multilingual support, real-time video processing, improved speech recognition accuracy, and personalized learning features to adapt content based on user preferences. Overall, the system provides a scalable and efficient solution for modern digital learning and content accessibility.

References

1. Alaa, T., Mongy, A., Bakr, A., Diab, M., & Gomaa, W. (2024). *Video Summarization Techniques: A Comprehensive Review.*
2. Mishra, P., Garg, K., & Rathi, N. (2023). *Video-to-Text Summarization Using NLP.*
3. Iswarya, M., Krishna, P. S., Naveen, K., Ganesh, M., & Yasin, M. (2023). *Video Transcript Summarization Using BERT.*
4. Ilampiray, P., Raju, N. D., Thilagavathy, A., et al. (2023). *Video Transcript Summarizer.*
5. Malu, R., Andhale, S., Potdar, V., & Khatavkar, T. S. (2023). *Video Summarization Using NLP.*
6. Shah, D., Namdev, U., Dedhia, M., Kanani, P., & Desai, R. (2022). *Video to Text Summarisation and Timestamp Generation.*
7. Yadav, S., Ekbal, A., & Bhattacharyya, P. (2021). *Automatic Generation of MCQs from Educational Texts in Hindi.*
8. Kolhe, S., Sharma, R., & Pise, S. (2020). *Video Summarization Using Highlight Detection and Deep Learning.*
9. Otani, M., Nakashima, Y., Rahtu, E., Heikkilä, J., & Yokoya, N. (2016). *Video Summarization Using Deep Semantic Features.*
10. Lin, C. Y. (2004). *ROUGE: A Package for Automatic Evaluation of Summaries.*
11. Vaswani, A., et al. (2017). *Attention Is All You Need.* NeurIPS.
12. Devlin, J., et al. (2019). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.* NAACL.
13. Zhang, J., Zhao, Y., & LeCun, Y. (2020). *Multimodal Fusion Techniques for Video Understanding.* IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI).
14. Gong, Y., & Liu, X. (2001). *Generic Text Summarization Using Relevance Measures.* SIGIR.
15. See, A., Liu, P. J., & Manning, C. D. (2017). *Get to the Point: Abstractive Summarization with Pointer-Generator Networks.* ACL.
16. Ekatpure, J. N., Tavate, C. S., Malshikare, S. S., Khomane, A. B., & Tamboli, M. J. L. (2025). Artificial intelligence based virtual keyboard and mouse for computer. *International Journal on Advanced Computer Theory and Engineering*, 14(1), 449–456.
17. Ekatpure, J. N., Mohite, S. D., Shinde, A. A., Shirkande, N. B., & Upase, V. V. (2025). Campus recruitment system using machine learning. *International Journal on Advanced Computer Theory and Engineering*, 14(1), 427–432.
18. Aware, D. B., Sayyad, S. R., Shaikh, A. H., Thombare, S. B., & Ekatpure, J. N. (2025). Translation assistant for converting sign language to text and audio. *International Journal on Advanced Computer Engineering and Communication Technology*, 14(1), 445–449.
19. Ekatpure, J. N., Aware, D. B., Shaikh, A. H., Sayyad, S. R., & Thombare, S. B. (2024). A comprehensive survey on sign language translation systems: Bridging gestures, text, and audio for enhanced communication. *International Journal of Recent Advances in Engineering and Technology*, 13(2), 15–21.
20. Ekatpure, J. N., Tavate, C., Malshikare, S., Khomane, A., & Tamboli, M. J. (2024). Advancements in AI-powered virtual keyboards and mice: A survey of cutting-edge technologies for modern computing. *International Journal on Advanced Computer Theory and Engineering*, 13(2), 52–57.