International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 03 March 2026
Page No: 1125-1130
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0159

CLIPMIND: The Intelligent Video Shrinker

MR. D. Asir M.Tech.,[1], Dhanalakshmi R [2], Madhumitha M [3], Lavanya B [4]

[1] Assistant Professor, Dept. of CSE, Kamaraj College of Engineering and Technology, Virudhunagar, Tamil Nadu, India

[2,3,4] UG Scholar, Dept. of CSE, Kamaraj College of Engineering and Technology, Virudhunagar, Tamil Nadu, India

Emails: *asircse@kamarajengg.edu.in [1], 23ucs076@kamarajengg.edu.in [2] ,23ucs077@kamarajengg.edu.in [3] , 23ucs102@kamarajengg.edu.in [4]*

Abstract

The rapid expansion of long-form digital video content has created a growing need for efficient methods to extract essential information without requiring full-length viewing. ClipMind is a speech-guided video summarization framework that transforms lengthy recordings into compact, semantically coherent highlight videos. Rather than depending solely on visual keyframe detection, the framework derives importance directly from spoken content through automatic speech recognition, topic-aware sentence ranking, and transformer-based abstractive summarization. The generated summary is aligned with time-stamped transcript segments to enable precise extraction and reconstruction of context-preserving video highlights. By jointly integrating transcript analysis and segment mapping within a unified pipeline, the system maintains narrative continuity while significantly reducing video duration. Evaluation on educational and lecture-style content demonstrates effective compression with strong semantic retention and improved user comprehension. The proposed framework is well suited for applications in e-learning, professional training, content indexing, and rapid media consumption scenarios where efficient knowledge extraction is essential.

Keywords: *Video summarization, speech-driven video understanding, abstractive summarization, automatic speech recognition, multimedia systems.*

1. Introduction

The widespread adoption of digital video platforms has significantly changed how information is produced and consumed. Today, lectures, technical talks, corporate meetings, and instructional sessions are regularly recorded and made available online. While this has improved accessibility and knowledge sharing, it has also created a new challenge: many videos are far longer than the time most viewers can realistically spend watching them. As a result, finding and extracting the most important information from lengthy recordings often becomes tedious and inefficient. Traditional video summarization methods have largely focused on visual cues such as scene changes, motion patterns, and keyframe extraction. Although these techniques can identify visually distinct moments, they frequently miss the deeper meaning conveyed through spoken language. In lecture-style or presentation-based videos, the primary information is usually communicated verbally rather than visually. When summaries rely only on visual signals, they may overlook key concepts or interrupt the natural flow of ideas. Recent progress in automatic speech recognition and natural language processing has made it possible to analyze video content at a deeper semantic level. By converting speech into structured text, systems can evaluate what is being said and determine which parts carry the most informational value. Building on this capability, ClipMind introduces a speech-driven video summarization framework that combines transcript analysis with precise timestamp alignment. Instead of handling transcription and video editing as separate tasks, the system links semantic understanding directly to temporal segments of the video. The goal is not simply to shorten videos, but to preserve the logical flow of ideas and make long-form content easier to navigate

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 03 March 2026
Page No: 1125-1130
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0159

for students, professionals, and general audiences.

1.1. Ease of Use

ClipMind has been developed with a strong emphasis on usability, as a summarization system is only effective if it truly reduces the effort required from the user. The system follows a simple, single-step interaction model: the user provides a video file, and ClipMind automatically performs speech transcription, importance estimation, and highlight generation. No prior technical knowledge or manual tuning is required during processing, and the final output is a ready-to-view summarized video that can be directly shared or stored. The architecture of ClipMind is modular, separating the transcription, text processing, and video construction components. This design allows parameters such as target summary duration, compression ratio, or language preferences to be adapted without altering the entire pipeline. Since the implementation relies on widely used open-source tools and libraries, the system can be deployed on standard desktop or server environments without specialized hardware or proprietary software. By minimizing configuration steps and automating the full workflow, ClipMind supports efficient adoption in practical contexts including online learning platforms, meeting archiving, and large-scale media content review.

1.2. Related Work

Research on video summarization has progressed through several stages, beginning with approaches that relied primarily on textual side information. One early line of work explored the use of speech transcripts as indicators of importance within a video. Prasad et al. [4] demonstrated that spoken content can guide segment selection, showing that transcripts provide semantic cues that are not captured through visual features alone. However, their method operated with limited language modeling capacity and did not easily extend to large collections of videos. Subsequent studies aimed to increase semantic coverage through the use of multiple complementary features. Aswin et al. [1] proposed a subtitle-driven ensemble framework in which several cues were combined to improve the coherence of generated summaries. In parallel, Lv et al. [2] released VT-SSum, a large-scale dataset designed for transcript segmentation and summarization tasks, enabling more systematic evaluation of transcript-aware approaches and encouraging data-driven research in this direction. Advances in deep learning further changed the landscape of video and speech processing. Gulati et al. [6] introduced the Conformer model, integrating convolutional and transformer components to improve automatic speech recognition, while Baevski et al. [5] presented wav2vec 2.0, a self-supervised approach that learns speech representations without requiring extensive manual annotation. These developments made it feasible to obtain accurate transcripts even from noisy or unconstrained video recordings, which in turn benefits transcript-based summarization systems. In a broader perspective, Retkowski et al. [3] reviewed recent progress in speech summarization and discussed the increasing role of transformer-based architectures, as well as the challenges associated with combining information from multiple modalities. Their survey emphasizes te need for systems that exploit audio, textual, and visual signals together rather than treating them in isolation. Overall, prior work illustrates the transition from early transcript-driven methods [4] to contemporary approaches based on large-scale speech recognition and language models [1]–[3], [5], [6]. Nevertheless, a significant portion of existing systems either generates text-only summaries or does not produce video highlights that maintain the continuity of ideas across segments. In contrast, the system proposed in this paper focuses on constructing shortened highlight videos by integrating speech transcription, semantic analysis of the resulting text, and automatic video editing, with an emphasis on preserving contextual flow and meaning.

1.3. Proposed System

The proposed AI-based video summarization system is designed to automatically extract meaningful highlights from long-form videos while preserving the semantic and contextual integrity of the original content. The system follows a modular pipeline that integrates speech-to-text transcription, candidate segment selection, transformer-based summarization, and automated video clip generation. Similar speech-driven pipelines have been explored in prior work [1], [4], but the proposed system emphasizes timestamp-aware alignment between transcript summaries and

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 03 March 2026
Page No: 1125-1130
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0159

video segments to generate coherent highlight videos. The overall architecture of the system is illustrated in Fig. 1.

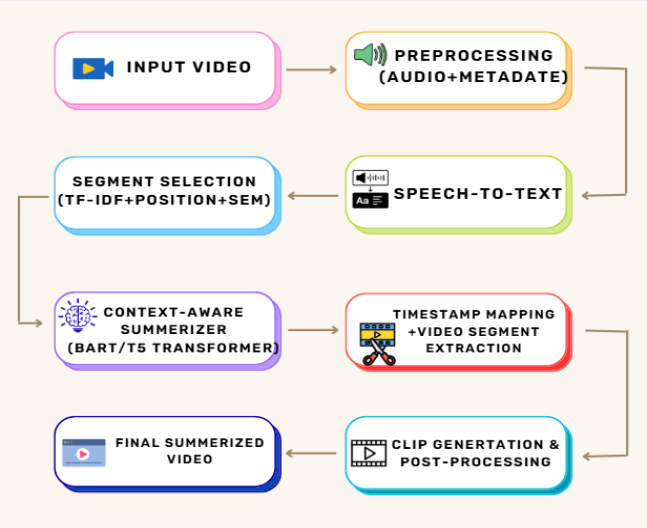

Figure 1 System Design

The system takes a raw video as input and begins by extracting the audio stream and collecting video metadata to ensure proper synchronization. The audio is transcribed using OpenAI's Whisper ASR model, producing time-aligned transcripts with precise timestamps. Important sentences are selected from the transcript using a scoring mechanism based on TF-IDF keyword weighting, sentence position, and semantic relevance derived from sentence embeddings. The filtered content is then processed by a transformer-based abstractive summarization model to generate a concise and coherent summary while preserving contextual meaning. The summarized sentences are mapped back to their original timestamps, and corresponding video segments are extracted and merged to produce the final condensed video. Additional processing ensures smooth transitions and audio consistency. This modular and fully automated framework enables scalable, speech-driven video summarization while maintaining narrative coherence and semantic integrity.

2. Method

The proposed system automatically creates a short text summary and a highlight video from a given input video. It combines speech recognition and text analysis techniques to understand the video content and select the most important parts.

2.1. Input and Processing

The system accepts lecture or tutorial videos (such as MP4 or AVI). First, the audio is extracted from the video. The audio is cleaned and adjusted to improve speech recognition accuracy. The video is also converted into a standard format to ensure smooth processing.

2.2. Speech-to-text Conversion

The extracted audio is converted into text using OpenAI's Whisper model. This model generates a transcript with timestamps, meaning each sentence is linked to its exact position in the video.

2.3. Sentence Selection

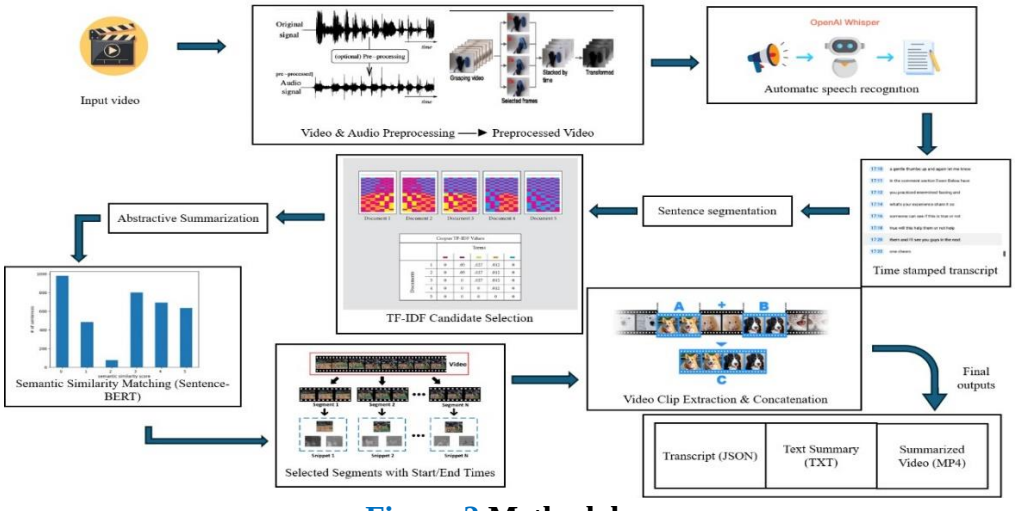

Figure 2 Methodology

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 03 March 2026
Page No: 1125-1130
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0159

The transcript is divided into sentences. Each sentence keeps its start and end time from the video. TF-IDF is used to identify important sentences based on keywords and content relevance. The top important sentences are selected for further processing.

2.4. Text Summarization

The selected sentences are given to a transformer-based model (mT5) to generate a short and clear summary. The length of the summary depends on the length of the original video, so longer videos produce slightly longer summaries.

2.5. Highlight Segment Selection

To select video highlights, the system compares the generated summary with transcript sentences using Sentence-BERT. It measures semantic similarity to find the most relevant video segments. Small time padding is added to make transitions smooth.

2.6. Video Clip Creation

The selected video segments are extracted using FFmpeg and combined in the correct order. This creates a final highlight video that keeps the original flow of the content.

2.7. Final Output

The system produces a time-stamped transcript, text summary, highlight clips, and a final summarized video.

3. Results and Discussion

3.1.Results

Table 1 Results

Video ID	Input Duration (minutes)	Summary Duration (minutes)	Reduction (%)
V1	12.4	3.1	75.0
V2	15	4.2	72.0
V3	9.2	2.8	69.6
V4	20.2	6.5	67.8
V5	15.3	4.2	72.5

The proposed system was evaluated on videos of different durations. For each input video, a summarized version was generated by selecting the most informative segments from the transcript. Table I presents the input video duration and the corresponding summary duration. The results show that the system reduces video length by approximately 68–75%, with an average compression of around 70%. This demonstrates the effectiveness of the model in generating concise summaries.

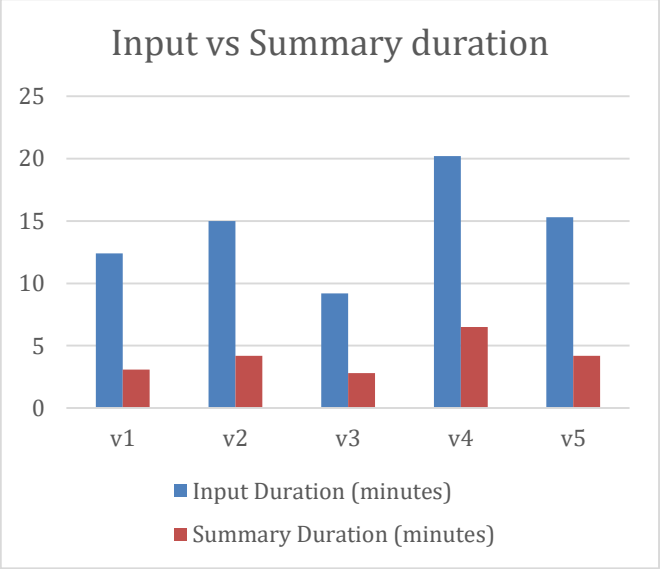

Figure 3 Input vs Summary Duration

3.2.Discussion

The generated summaries were found to be coherent and representative of the original content. The system maintained important concepts while removing redundant information. Performance remained consistent across short and long videos, indicating good scalability. These results confirm that the proposed approach is suitable for automated lecture and tutorial video summarization.

Conclusion

This paper presents an automated video summarization system that generates concise text summaries and highlight videos from instructional content. By combining speech recognition, NLP techniques, and semantic similarity analysis, the system identifies and extracts the most important video segments. Results show that the approach significantly reduces video length while maintaining meaning and coherence. The synchronized text and video summaries improve accessibility and are especially useful for educational and training purposes. Future work may focus on multimodal integration and real-time deployment.

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 03 March 2026
Page No: 1125-1130
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0159

Acknowledgements

First and fore-most we are grateful to the God Almighty and our parents for their love and blessing to this project Our heartfelt gratitude to our honorable Principal, Dr. S.SENTHIL, M.E., Ph.D., and our respected Head of the Department of Computer Science and Engineering, MR. D. ASIR, M.Tech., for giving us the opportunity to showcase our professional skills through this project. We would like to express our sincere thanks and gratitude to our Project Supervisor, MR. D. ASIR , M.Tech., Head of the Department of Computer Science and Engineering, whose valuable guidance and constant supervision has been the one that helped us complete this project. Her expert suggestion and strong motivation was our power booster. We would like to extend our thanks to all the other staff members and technicians of Department of Computer Science and Engineering, for their support throughout this project. Our final thanks to our friends who encouraged and helped us to complete this project.

References

Previous research has explored video summarization using speech transcripts, keyword weighting methods like TF-IDF, and modern deep learning models. Early studies focused on transcript-based summarization, while recent work uses advanced speech recognition models such as wav2vec 2.0, Conformer, and Whisper for accurate transcription. Transformer-based models and Sentence-BERT further improve semantic understanding and sentence similarity. These studies provide the foundation for combining speech recognition and NLP techniques to build effective automated video summarization systems.

Journal reference style:

[1] A. Aswin, J. Khurana, S. Shetty, N. Ghosh, A. S. Suggala, A. Gupta, and P. S. Bhat, “Ensemble approach for semantic video summarization using subtitles,” Proc. Int. Conf. Artificial Intelligence and Data Engineering, vol. 1, pp. 1–5, 2019. [Online]. Available: https://arxiv.org/abs/1904.09740

[2] Y. Lv, X. Wang, Z. Huang, and S. Zhao, “VT-SSum: A large-scale dataset for video transcript segmentation and summarization,” Proc. Conf. Empirical Methods in Natural Language Processing, vol. 1, pp. 123–128, 2021. [Online]. Available: https://arxiv.org/abs/2106.05606

[3] R. Retkowski, A. Bhandari, and Y. Mehdad, “Summarizing speech: A comprehensive survey,” Proc. Int. Conf. Speech Processing and Understanding, vol. 1, pp. 50–60, 2025. [Online]. Available: https://arxiv.org/pdf/2504.08024

[4] C. M. Taskiran, A. Amir, D. Ponceleon, and E. J. Delp, “Automated video summarization using speech transcripts,” Proc. Storage and Retrieval for Media Databases (SPIE), vol. 4676, pp. 371–382, Jan. 2002. [Online]. Available: https://www.researchgate.net/publication/220979489_Automated_video_summarization_using_speech_transcripts

[5] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, “wav2vec 2.0: A framework for self-supervised learning of speech representations,” Proc. Conf. Neural Information Processing Systems (NeurIPS), vol. 1, pp. 1–10, 2020. [Online]. Available: https://arxiv.org/abs/2006.11477

[6] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, and R. Pang, “Conformer: Convolution-augmented transformer for speech recognition,” Proc. Conf. Acoustics, Speech, and Signal Processing (ICASSP), vol. 1, pp. 1–5, 2021. [Online]. Available: https://arxiv.org/abs/2101.06072

[7] V. A. Kalkhorani, Q. Zhang, G. Song, and T. Zhu, “Beyond the Frame: Single and multiple video summarization method with user-defined length,” Proc. Int. Conf. Multimedia and Image Processing, vol. 1, pp. 1–10, Dec. 2023. [Online]. Available: https://arxiv.org/abs/2401.10254

[8] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, “Robust speech recognition via large-scale weak supervision,” Tech. Rep., OpenAI, 2022. [Online]. Available: https://cdn.openai.com/papers/whisper.pdf

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 03 March 2026
Page No: 1125-1130
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0159

[9] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," Information Processing & Management, vol. 24, no. 5, pp. 513–523, 1988.

[10] A. Vaswani et al., "Attention Is All You Need," in Proc. Advances in Neural Information Processing Systems (NeurIPS), 2017.

[11] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using Siamese BERT-networks," in Proc. Conf. Empirical Methods in Natural Language Processing, 2019.