AI & ML Enabled Video Analysis and Interpretation

Vivek Chauhan[1], Vivek Sharma[2], Yash Rajput[3], Shani Rathore[4], Mr. Suman Kumar Jha[5], Badal Bhushan[6]

[1,2,3,4]B. Tech (CSE) -Final Year Student, Dept Computer Science & Engineering, IIMT College of Engineering, Greater Noida

[5,6]Project Supervisor, Dept. of Computer Science & Engineering, IIMT College of Engineering, Greater Noida,, Greater Noida, UP, India

DOI: https://doi.org/10.51584/IJRIAS.2025.10120067

Received: 24 December 2025; Accepted: 29 December 2025; Published: 16 January 2026

ABSTRACT

With video content absolutely everywhere these days—on learning platforms, in business settings, across social media—trying to analyze it all by hand has become practically impossible. Our paper describes a framework we built that uses AI and machine learning to make understanding videos much simpler, whether you're uploading your own footage or just sharing a link to something online.

Here's how it works: the system examines what's actually happening on screen while also listening to the audio, then brings everything together into summaries that actually make sense. We're using a Transformer-based model that's really good at figuring out how different moments in a video relate to each other and what they mean in context. After you get your summary, there's also a lightweight language model that lets you have an actual conversation about what you watched—you can ask questions and get answers that show a real understanding of the content.

One thing that sets our approach apart is that we focus on videos that have already been recorded rather than trying to analyze everything in real-time. This decision makes the whole system more flexible and honestly much easier to implement in different scenarios. When we tested it on all kinds of video samples, it consistently produced summaries that held together well and answered questions in ways that showed it understood the context. We think this could be really valuable for education, content analysis, and helping organizations manage their knowledge better.

Keywords: Video Analysis, Video Summarization, Artificial Intelligence, Machine Learning, Transformer Models, Multimodal Learning, Video Processing, Video Interpretation

INTRODUCTION

We're living in an era where video content is everywhere—YouTube, online courses, company training videos, you name it. This explosion of video has created a real problem: how do we quickly understand and find what we need in all these videos? Unlike reading a document, videos are tricky to work with. They mix together images, sounds, speech, and text, all at the same time. Going through hours of video manually to find information? That's just not practical anymore.

This is where AI and smart computer programs come to the rescue. These technologies have changed the game for video analysis, making it possible to automatically create summaries, detect different scenes, generate captions, and figure out what's actually happening in a video. In the past, video processing tools used basic, hand-coded rules. The problem was, these old methods couldn't really understand the deeper meaning or follow how things changed throughout a video. Thankfully, recent advances in AI—particularly newer learning models—have dramatically improved how computers can understand patterns and meaning in videos.

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume X Issue XII December 2025

Here's an interesting observation: most existing video tools are built for live scenarios like security cameras or live streaming. But in reality, many everyday uses involve pre-recorded videos. Think about it—people often want to upload a video file or share a link and get insights from it afterwards. These situations need more than just a quick summary. When someone's trying to understand a long lecture or a recorded meeting, they often want to dig deeper and ask specific questions about what happened.

That's exactly what we're solving in this paper. We present a practical system that uses AI to analyze pre-recorded videos in a smarter way. Our system does two main things: first, it automatically creates a text summary of what's in the video; second, it lets users chat and ask questions about the content—almost like having a conversation about the video. We built this using modern AI models that convert video to text, and then answer user questions. The nice thing about our design is that it works in steps: the summary part feeds directly into the question-answering part, creating a smooth flow.

We believe this work makes three important contributions:

1. We've designed a complete system for analyzing pre-recorded videos using straightforward AI models that others can actually use and build upon.

2. We've connected video summarization with question-answering, giving users a much better way to understand video content beyond just reading a simple summary.

3. We've demonstrated an approach that's both practical and scalable, making it useful for real-world situations in education, content analysis, and information management.

The rest of this paper walks through our work step by step. Section II reviews what others have done in video summarization and understanding. Section III explains our method and how we built the system. Section IV covers our experiments and testing. Section V discusses what we found and the challenges we faced. Finally, Section VI wraps up with conclusions and ideas for future work.

Related Work

Video analysis and interpretation has been widely explored in areas like video summarization, working with multiple types of data, and understanding video content. Early work mainly dealt with pulling out key frames and identifying scene changes, using basic visual elements like color patterns, movement tracking, and grouping similar frames together to create shortened video versions [14]–[17]. While these methods were fast and didn't require much computing power, they couldn't really understand what was happening in the videos and often missed important context.

As deep learning became more popular, video summarization techniques using neural networks showed much better results. Combining different types of neural networks made it possible to track changes over time in videos and better identify which frames were most important [1], [11]. Special focusing mechanisms made things even better by helping models zoom in on the most relevant parts of a video [12].

More recent work has looked at bringing together multiple types of information—visuals, audio, and text—to create more meaningful video summaries. Some models use layered structures that work at the shot level, combining different types of features to understand connections across longer video segments [3]. Meanwhile, newer approaches use special attention techniques to keep the story coherent across frames [5], [6]. Researchers have also tackled the challenge of keeping different types of information consistent with each other, using various techniques to bridge gaps between different data types [4].

Modern AI models have opened up new possibilities for understanding videos through detailed video captioning and breaking down events within videos, allowing for much more detailed interpretation [8]. On top of this, automatic learning methods for audio processing, including speech recognition models, have improved how well we can convert spoken words to text, making text-based summaries and analysis more accurate [9].

Even with all this progress, most current methods focus on creating summaries or analyzing videos in isolation, without much consideration for interactive, user-driven exploration. Some systems have tried to address this by generating different types of summaries and allowing users to search based on specific questions [13], [19]. However, combining video summarization with chat-like interfaces where users can ask questions and get answers remains largely unexplored.

This work builds on previous research by introducing an AI and machine learning–powered video analysis system that brings together advanced summarization with interactive exploration. This lets users explore video content by simply asking questions in everyday language.

PROPOSED METHODOLOGY

This section presents the proposed AI and ML enabled framework for video analysis and interpretation. The methodology is designed as a modular pipeline, where the output of one model serves as the input to the next, enabling progressive refinement of information extracted from user-provided videos.

A. Input Acquisition and Preprocessing

The system accepts video input in two formats:

(i) a raw uploaded video file, or

(ii) a video URL (e.g., YouTube link).

For URL-based input, the video is first downloaded and converted into a standard format. The video is then decomposed into frames at fixed intervals, while the audio stream is extracted separately. Basic preprocessing operations such as frame resizing, normalization, and audio sampling are applied to ensure compatibility with downstream models.

B. Video Encoding Using Vision Transformer

To capture the visual semantics of the video, a Vision Transformer (ViT)–based encoder is employed. Unlike traditional convolutional models that focus on local patterns, the transformer encoder processes the video frames as a sequence of visual tokens, enabling the model to capture global contextual relationships across frames.

Each frame is divided into patches, embedded into a fixed-dimensional representation, and passed through multiple self-attention layers. The resulting encoded representations capture objects, scenes, and temporal continuity present in the video. These visual embeddings form a compact semantic representation of the video content.

C. Textual Summary Generation Using Transformer-Based Language Model

The visual embeddings generated by the video encoder are passed as input to a Transformer-based text generation model. This model is responsible for producing a natural language summary of the video.

By attending to the encoded visual information, the language model generates a coherent textual description that highlights the main events, themes, and key transitions within the video. The summarization process is performed in an manner, ensuring that the entire video context is available during generation, which improves summary quality and completeness.

D. Knowledge Extraction and Representation

The generated summary, along with intermediate semantic cues (such as key segments and timestamps), is further processed to extract important entities, actions, and topics. This information is stored in a structured textual form that acts as a knowledge source for interpretation.

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume X Issue XII December 2025

Instead of relying on complex symbolic reasoning, the system uses lightweight text embeddings to represent the extracted knowledge, allowing efficient retrieval and reuse in subsequent stages.

E. Conversational Interpretation Module

To enable interactive understanding, the extracted textual knowledge is provided as input to a chatbot-based interpretation module. When a user asks a question related to the video, the chatbot retrieves relevant information from the stored summary and generates a context-aware response.

This approach allows users to explore the video content conversationally, such as asking for clarifications, explanations of specific segments, or high-level insights, without reprocessing the video repeatedly.

F. Overall Workflow

The complete methodology follows a sequential pipeline:

1. Video input acquisition and preprocessing
2. Visual feature encoding using a Vision Transformer
3. Textual summary generation using a Transformer-based language model
4. Knowledge extraction and storage
5. Conversational interpretation through a chatbot interface

By separating video understanding and interaction into distinct yet connected stages, the proposed framework achieves efficient, scalable, and interpretable video analysis.

System Architecture

The proposed system architecture is designed to enable video analysis and interpretation using a modular and scalable AI-based pipeline. The architecture supports both user-uploaded video files and video URLs as input and processes them in a structured manner to generate meaningful summaries and interactive interpretations. The overall design follows a pipeline-based approach, where the output of one module serves as the input to the next, ensuring refined and context-aware results.

A. Input Acquisition Layer

The system begins with an input acquisition layer that accepts:

- Pre-recorded video files (e.g., MP4, AVI), or
- Online video URLs (such as YouTube links).

In the case of URLs, the video is first downloaded and converted into a standard format. This layer ensures uniformity in video resolution, frame rate, and encoding, which simplifies downstream processing.

B. Video Preprocessing Layer

Once the video is obtained, preprocessing is performed to prepare the data for analysis. This includes:

- Frame extraction at fixed intervals,
- Audio stream separation,
- Shot and scene boundary detection.

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume X Issue XII December 2025

This step reduces redundancy, filters irrelevant frames, and segments the video into meaningful temporal units, enabling efficient analysis without requiring real-time processing.

C. Visual Feature Encoding Module

The extracted frames are passed to a vision-based transformer or lightweight CNN-based encoder, which converts raw visual information into high-level feature representations. Instead of relying on low-level pixel values, the encoder captures semantic patterns such as objects, scenes, and actions. These embeddings provide a compact yet informative representation of the visual content.

D. Audio and Text Processing Module

Parallel to visual encoding, the audio stream is processed using a basic speech-to-text model. The generated transcript captures spoken information, narration, or dialogue present in the video. This textual data is further cleaned and segmented to align with the corresponding video timestamps, enabling multimodal synchronization between visual and audio information.

E. Multimodal Fusion and Representation Layer

In this layer, visual embeddings and textual representations are combined using a simple multimodal fusion strategy. The objective is to align semantic information across modalities rather than performing complex cross-attention mechanisms. The fused representation serves as a unified knowledge source that reflects both what is shown and what is spoken in the video.

F. Video Summarization Module

The unified multimodal representation is fed into a transformer-based text generation model responsible for producing a concise and coherent video summary. The summarization module focuses on extracting key events, important transitions, and dominant themes while preserving the overall context of the video. The generated summary is textual and timestamp-aware, making it suitable for further interpretation.

G. Interpretation and Knowledge Indexing Layer

The extracted summary and intermediate semantic representations are stored in an indexed format. This layer acts as a lightweight knowledge base that enables structured access to video content. The indexed information includes:

- Summary segments,
- Key timestamps,
- Scene-level descriptions.

H. Output Layer

The system produces two primary outputs:

1. An automatically generated textual summary of the provided video.
2. An interactive conversational interface that allows users to query and interpret the video content.

I. Architectural Advantages

The proposed architecture:

- Supports processing, avoiding real-time constraints.

- Uses simple and easily implementable models, making it suitable for academic environments.
- Ensures modularity, allowing individual components to be upgraded independently.
- Provides explainable and user-centric video interpretation.

Video Analysis and Interpretation Module

A. Video Analysis and Interpretation Module

The Video Analysis and Interpretation Module is the core component of the proposed system, responsible for transforming raw video input into meaningful semantic representations that can be understood, summarized, and interactively queried. Unlike low-level video processing techniques that focus only on frames or motion, this module emphasizes high-level semantic understanding of video content.

B.Video Analysis Phase

In the analysis phase, the input video—either uploaded directly by the user or obtained via a video URL—is first decomposed into a sequence of frames and audio segments. Visual information is extracted at regular intervals to capture scene transitions, object presence, and contextual cues. At the same time, the audio stream is processed to obtain textual transcripts using a speech-to-text model. These visual and audio features together provide a comprehensive representation of the video content.

To preserve temporal coherence, the video is segmented into logical units such as scenes or shots. This segmentation enables the system to understand the progression of events and maintain contextual continuity across the video. The extracted features are encoded into compact vector representations that summarize both spatial and temporal characteristics of the video.

C. Interpretation Phase

The interpretation phase focuses on converting the analyzed video features into human-understandable information. The encoded multimodal representations are passed to a transformer-based language model, which generates a concise textual summary describing the key events, themes, and conclusions of the video. This summary serves as a high-level abstraction of the video content.

In addition to summary generation, the interpretation module supports interactive understanding through a conversational interface. The extracted semantic representations are indexed and stored in an embedding space, enabling efficient retrieval of relevant video segments. When a user poses a query, the system retrieves the most relevant contextual information from the analyzed video and generates an appropriate response using a language model. This allows users to ask detailed questions such as explanations of specific scenes, clarification of concepts, or identification of important moments within the video.

EXPERIMENTAL SETUP AND RESULTS

This section describes the experimental configuration used to evaluate the proposed AI and ML enabled video analysis and interpretation system, along with the results obtained from multiple test cases.

A. Experimental Setup

To evaluate the effectiveness of the proposed framework, experiments were conducted on user-provided video content, including both uploaded video files and video URLs sourced from publicly available platforms. The selected videos varied in duration, content type, and complexity to ensure robustness of the system across diverse scenarios.

The video analysis pipeline processes each input video by extracting visual frames at fixed intervals and generating audio transcripts using a speech-to-text model. Visual features are encoded using a lightweight

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume X Issue XII December 2025

Transformer-based encoder, while textual features are represented using pre-trained text embeddings. These multimodal representations are then used for video summarization and interpretation.

The system was implemented using Python, with deep learning models executed on a standard computing environment. The summarization and interpretation modules operate sequentially, where the output of the video summarization model is passed as input to the interpretation and conversational module.

B. Evaluation Metrics

The performance of the proposed system was evaluated using both quantitative and qualitative metrics:

- ROUGE Scores (ROUGE-1, ROUGE-2) to measure the quality of generated summaries
- Semantic relevance score to evaluate alignment between video content and generated interpretation
- Response accuracy for chatbot-based question answering
- Processing time to assess computational efficiency

These metrics provide a balanced evaluation of both content quality and system performance.

C. Results and Analysis

Table I presents the summarization and interpretation performance of the proposed system across different video inputs.

Table I. Performance Evaluation of Video Analysis and Interpretation System

Video Type	Duration (min)	ROUGE-1	ROUGE-2	Accuracy
Education	**15**	**0.52**	**0.31**	**87**
News	**10**	**0.48**	**0.29**	**84**
Tutorial	**12**	**0.50**	**0.30**	**86**
Random	**8**	**0.55**	**0.34**	**89**

The results indicate that the proposed system generates contextually coherent and concise summaries, with higher accuracy observed in structured video content such as tutorials and lectures. The interpretation module demonstrates strong semantic understanding, enabling effective conversational interaction with the video content.

DISCUSSION

The experimental results demonstrate that integrating AI and ML models for video summarization and interpretation significantly enhances video understanding. The sequential model pipeline enables refined output generation, where summarized content improves the relevance and accuracy of conversational responses. While the system performs well for structured videos, performance slightly degrades for videos with noisy audio or rapid scene changes, indicating areas for future improvement.

Limitation And Future Work

Despite the effectiveness of the proposed AI and ML enabled video analysis and interpretation system, certain limitations remain that provide opportunities for future enhancements.

Limitations

1. **Dependence on Pretrained Models:**

The system relies on pretrained transformer-based models for video summarization and text generation. While this ensures ease of implementation and reduced training cost, it may limit adaptability to highly domain-specific videos such as medical or legal content.

2. **Computational Overhead:**

Transformer-based video encoding and multimodal fusion require significant computational resources, especially for long-duration or high-resolution videos. This may affect scalability when processing large video datasets.

3. **Evaluation Constraints:**

The experimental evaluation primarily focuses on summarization quality and chatbot response relevance. Human-centered evaluation, such as long-term user engagement and interpretability assessment, is limited..

Future Work

1. **Domain-Specific Model Adaptation:**

Future work can incorporate domain-adaptive fine-tuning techniques to improve performance in specialized areas such as education, healthcare, and legal video analysis.

2. **Enhanced Temporal Modeling:**

Integrating advanced temporal reasoning mechanisms, such as temporal transformers or graph-based event modeling, can further improve event-level interpretation and timeline-based querying.

3. **Interactive Visualization Support:**

Future versions may include visual explanation modules that display keyframes, timelines, and attention heatmaps to improve transparency and user trust in the system's outputs.

4. **Scalable Deployment:**

Optimizing the pipeline for distributed and cloud-based environments can enable real-time responsiveness and large-scale deployment for enterprise applications.

CONCLUSION

This paper presented an AI and ML enabled framework for video analysis and interpretation that focuses on understanding user-provided video content through automated summarization and interactive interpretation. By leveraging transformer-based models for video-to-text understanding and language-based interpretation, the proposed system is capable of extracting meaningful information from both raw videos and video URLs, and presenting it in a concise, user-friendly manner.

Unlike traditional video analysis approaches that emphasize low-level visual processing or real-time analysis, this work demonstrates how semantic understanding of video content can be achieved through a modular pipeline where the output of one model serves as structured input to another. The integration of video summarization with a conversational interpretation module enables users to not only consume condensed video content but also interactively query and explore the underlying information.

Experimental results indicate that the proposed approach effectively captures key events, maintains contextual coherence in summaries, and provides relevant responses to user queries. The modular nature of the system

ensures flexibility, ease of implementation, and adaptability to different video domains such as education, media analysis, and content review.

Overall, this research highlights the potential of transformer-based AI models in advancing video analysis from passive summarization to active interpretation, paving the way for more intelligent and user-centric video understanding systems.

REFERENCES

1. E. Apostolidis, E. Adamantidou, A. I. Metsai, V. Mezaris, and I. Patras, "Video summarization using deep neural networks: A survey," IEEE Transactions on Neural Networks and Learning Systems, 2024.
2. G. Peronikolis and C. Panagiotakis, "Personalized video summarization: A comprehensive survey of methods and datasets," Machine Learning and Knowledge Extraction, MDPI, 2024.
3. X. Xu, et al., "MHSCNet: Multimodal hierarchical shot-aware convolutional network for video summarization," arXiv preprint, 2024.
4. Y. Qiu, et al., "Semantics-consistent cross-domain video summarization via optimal transport alignment," arXiv preprint, 2023.
5. J. Park, et al., "Multimodal frame-scoring transformer for video summarization," arXiv preprint, 2023.
6. M. Krubiński and P. Pecina, "MLASK: Multimodal summarization of video-based news articles," in Proceedings of the European Chapter of the Association for Computational Linguistics (EACL), 2023.
7. M. Alaa, et al., "Video summarization techniques: A comprehensive review," SciTePress, 2022.
8. H. Zhou, et al., "End-to-end dense video captioning with masked transformer," IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022.
9. A. Baevski, et al., "wav2vec 2.0: A framework for self-supervised learning of speech representations," in Proceedings of the Neural Information Processing Systems (NeurIPS), 2021.
10. H. Ji, D. Hooshyar, et al., "A semantic-based video scene segmentation using a deep neural network," SAGE Journals, 2021.
11. Y. Otani, et al., "Video summarization using deep semantic features," IEEE Transactions on Circuits and Systems for Video Technology, 2021.
12. J. Zhong, et al., "Video summarization with attention-based encoder–decoder networks," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2020.
13. S. Sahoo, et al., "A unified multi-faceted video summarization system," arXiv preprint, 2020.
14. V. D. Desai, "A review paper on keyframe extraction techniques for video summarization," International Journal of Research and Analytical Reviews (IJRAR), 2019.
15. "Review of keyframe extraction techniques for video summarization," International Journal of Computer Applications (IJCA), 2019.
16. "Digital video summarization techniques: A survey," International Journal of Engineering Research and Technology (IJERT), 2018.
17. N. Ejaz, et al., "Adaptive key frame extraction for video summarization," Elsevier Journal, 2018.
18. H. Yu, et al., "Video summarization using U-shaped non-local network," Elsevier Journal, 2017.
19. Y. Zhu, et al., "Topic-aware video summarization using multimodal feature learning," Elsevier Journal, 2016.
20. S. Sahoo, et al., "A unified multi-faceted video summarization system," arXiv preprint, 2012.