

VideoIQ: A Multimodal Video SEO Intelligence System

1st E. Srinath

CSE Assistant Professor

Keshav Memorial Institute of Technology

Hyderabad 500029, India

esrinath77@gmail.com

2nd J. L. John Nicholas

IV year CSE Student

Keshav Memorial Institute of Technology

Hyderabad 500029, India

jljohnnicholas@gmail.com

3rd C. H. Venkata Sri Sai

IV year CSE Student

Keshav Memorial Institute of Technology

Hyderabad 500029, India

venkatsrisai670@gmail.com

4th S. Sai Ridhi

IV year CSE Student

Keshav Memorial Institute of Technology

Hyderabad 500029, India

ridhisriramula@gmail.com

5th J. Shruthi

IV year CSE Student

Keshav Memorial Institute of Technology

Hyderabad 500029, India

jerubandishruthee@gmail.com

Abstract—With the exponential growth of video-centric platforms, content visibility is no longer determined solely by creative quality but by how effectively a video aligns with search and recommendation mechanisms. Most creators rely on manual trial-and-error methods for Search Engine Optimization (SEO), often without understanding how their actual video content influences discoverability.

This project presents an end-to-end multimodal Video SEO Intelligence Platform that performs automated SEO optimization by directly interpreting the underlying audio-visual content of videos. The system extracts semantic information through visual frame analysis, speech-based transcription, and transcript-level natural language processing. These multimodal signals are fused to identify core themes and contextual keywords that accurately represent the video's intent.

The extracted content features are further refined using a trend-aware mapping mechanism that aligns keywords with platform-specific and time-sensitive search behaviour. The system generates optimized video descriptions and hashtags, assigns a quantitative SEO effectiveness score, and provides an estimated performance outlook prior to publishing.

Index Terms—Video processing, SEO, speech-to-text, generative AI, content analysis

I. INTRODUCTION

The exponential growth of video-centric platforms such as YouTube, Instagram, and TikTok has fundamentally transformed digital content consumption patterns. The increasing volume of video uploads has resulted in a highly competitive ecosystem, where content discoverability depends not only on content quality but also on effective Search Engine Optimization (SEO) strategies [31].

Traditional SEO workflows rely heavily on manual intervention, requiring content creators to independently define titles, descriptions, and metadata tags. This process often lacks systematic analysis of the underlying video content, leading to a disconnect between semantic meaning and metadata representation. Such inconsistencies negatively impact search ranking and audience engagement [4].

Recent advancements in automatic speech recognition (ASR) and natural language processing (NLP) have enabled

automated extraction of meaningful information from multimedia data. Techniques for audio feature extraction and speech understanding have significantly improved transcription accuracy [28], while text analysis and information extraction methods have enhanced the ability to derive structured insights from unstructured content [1]. Furthermore, feature ranking and keyword extraction approaches have been widely adopted for identifying relevant content attributes [2].

Despite these advancements, existing systems primarily operate on manually provided textual inputs or address isolated subtasks such as keyword extraction, tagging, or summarization. There is a lack of unified systems capable of directly analyzing raw video content and generating structured SEO metadata in an end-to-end manner.

To address this limitation, this paper proposes **VideoIQ**, a multimodal video SEO intelligence system that: (i) automatically extracts speech from video, (ii) generates a validated transcript, (iii) produces optimized SEO outputs through AI-driven content analysis, and (iv) estimates engagement potential using a content-driven trend analysis module. The primary contributions of this work are:

- A fully automated, end-to-end pipeline that transforms raw video input into structured SEO metadata.
- A transcript quality validation mechanism to ensure reliable downstream processing.
- A lightweight trend scoring model based on semantic and lexical features.
- Experimental evaluation across multiple video categories demonstrating system effectiveness.

II. RELATED WORK

A. Automatic Speech Recognition

Automatic speech recognition (ASR) has significantly improved with the advancement of deep learning-based audio processing techniques. Recent work in audio feature extraction and speech understanding has enabled robust transcription across diverse acoustic conditions [28]. Additionally, deep

neural network-based audio models have demonstrated improved performance in extracting meaningful representations from speech signals [25]. These developments provide the foundation for converting video audio streams into structured textual data.

B. Natural Language Processing for Content Generation

Natural language processing techniques have been widely used for information extraction, summarization, and keyword generation. Methods for extracting structured information from textual data have shown strong effectiveness in deriving meaningful insights from unstructured content [1]. Furthermore, feature ranking and classification techniques have improved the identification of relevant keywords and semantic features [2]. These approaches are critical for generating SEO metadata from transcript data.

C. SEO and Content Optimization

SEO systems aim to improve content visibility through keyword optimization and ranking strategies. Existing approaches focus on enhancing ranking performance using embedding-based models and metadata optimization techniques [4]. Additionally, research in metadata extraction and keyword mapping has demonstrated the importance of structured content representation for improving discoverability [17]. However, these systems typically rely on manually provided inputs and do not directly analyze video content.

D. Video Tagging and Trend Analysis

Video tagging and recommendation systems have leveraged multimodal learning approaches to combine audio, visual, and textual features [5]. Hashtag recommendation techniques for micro-videos have been explored using content-based and semantic approaches [6]. Furthermore, studies in video recommendation highlight the importance of metadata and tagging in improving content visibility and engagement [31]. Despite these advancements, most existing systems address individual components such as tagging, recommendation, or summarization in isolation. The proposed system extends prior work by integrating transcription, SEO generation, and trend analysis into a unified pipeline operating directly on raw video input.

III. SYSTEM ARCHITECTURE

VideoIQ follows a three-tier client-server architecture comprising a frontend presentation layer, a backend processing layer, and an AI inference layer. The overall architecture is illustrated in Fig. ??.

The system is designed to enable seamless integration of video processing, transcription, and SEO generation within a unified pipeline.

A. Frontend Layer

The frontend is implemented using React.js and provides an interactive interface for user interaction. It supports:

- A video upload interface with file validation and size constraints

- Real-time processing feedback through asynchronous communication with the backend
- Structured visualization of outputs including summaries, keywords, hashtags, and trend analytics

The frontend communicates with the backend through RESTful APIs and dynamically updates the user interface based on processing status and results.

B. Backend Layer

The backend is implemented using Flask and serves as the central orchestration unit of the system. It is responsible for:

- Handling multipart video file uploads and temporary storage
- Executing the sequential processing pipeline
- Managing API communication with AI services
- Structuring and returning responses to the frontend

The backend ensures robustness by incorporating validation checks and error handling at each stage of processing.

C. AI Processing Layer

The AI processing layer consists of three primary modules:

- **Speech-to-Text Module:** Converts extracted audio into textual transcripts using deep learning-based speech recognition techniques [28].
- **SEO Generation Module:** Applies natural language processing and information extraction techniques to generate summaries, keywords, and metadata [1], [2].
- **Trend Analysis Module:** Performs feature-based evaluation of transcript content to estimate engagement potential and identify relevant trends, inspired by multimodal tagging and recommendation approaches [5], [31].

D. System Characteristics

The architecture is designed with the following properties:

- **Modularity:** Each component operates independently, allowing easy updates or replacement.
- **Scalability:** The system can handle varying video sizes and workloads.
- **Extensibility:** Additional modules such as visual analysis or multilingual support can be integrated.
- **Robustness:** Error handling mechanisms ensure stable operation under edge cases.

The modular and layered design enables efficient processing while maintaining flexibility for future enhancements.

IV. METHODOLOGY

VideoIQ processes video input through a structured multi-stage pipeline that transforms raw video data into SEO-relevant outputs. Each stage of the pipeline is formally described below.

A. Audio Extraction

Raw video input V is processed to extract the audio stream using a multimedia processing framework. The extraction operation is defined as:

$$A(t) = E(V(t)) \quad (1)$$

where $V(t)$ represents the video signal at time t , $A(t)$ denotes the extracted audio waveform, and $E(\cdot)$ represents the extraction operator. The audio is normalized to a consistent format (mono-channel, 16 kHz sampling rate) to ensure compatibility with downstream processing modules.

B. Speech-to-Text Conversion

The extracted audio signal $A(t)$ is converted into a textual transcript using deep learning-based speech recognition techniques:

$$T = \text{STT}(A; \vartheta) \quad (2)$$

where T represents the generated transcript and ϑ denotes model parameters. Recent advances in audio feature extraction and speech processing have significantly improved transcription accuracy across diverse environments [28], [25].

C. Transcript Quality Validation

To ensure reliability, the transcript T is evaluated using a quality metric Q :

$$Q(T) = \frac{|W(T)|}{|K(T)|} \quad (3)$$

where $|W(T)|$ is the total word count and $|K(T)|$ is the number of unique content-bearing tokens. Transcripts satisfying $Q(T) \geq \tau$ are considered valid, where τ is a predefined threshold. This step filters out low-information transcripts resulting from noise or non-speech content.

D. SEO Content Generation

The validated transcript T is processed using natural language processing techniques to generate structured SEO outputs:

$$S = G(T, P) \quad (4)$$

where $S = \{s, K, H, \hat{t}\}$ represents the generated summary s , keywords K , hashtags H , and title $\hat{t}$, and P denotes the structured prompt or instruction set. Information extraction and feature ranking methods are employed to identify relevant textual features [1], [2]. Additionally, keyword extraction techniques play a significant role in identifying high-impact SEO terms [19].

E. Content-Driven Trend Analysis

A trend score Φ is computed using transcript-derived semantic features:

$$\Phi = \alpha L + \beta D + \gamma K_f \quad (5)$$

where:

- $L = \log(1+N_w)$ represents the normalized content length
- $D = \frac{|V_w|}{N_w}$ denotes lexical diversity

- K_f represents keyword frequency relevance

The coefficients α, β, γ are weighting parameters such that $\alpha + \beta + \gamma = 1$. This formulation captures content richness and semantic diversity as indicators of engagement potential. Similar approaches in multimodal tagging and recommendation systems have demonstrated the importance of content features in predicting user engagement [5], [31].

F. Output Structuring

The final outputs are structured into a standardized format for frontend visualization. The results include summary, keywords, hashtags, title suggestions, and trend analytics, enabling users to directly utilize the generated SEO insights.

V. SYSTEM WORKFLOW

The overall workflow of the VideoIQ system follows a structured and sequential pipeline that transforms raw video input into meaningful and actionable SEO insights. The workflow is illustrated in Fig. ???. Each stage in the pipeline is designed to progressively process and refine the input data.

- 1) **Video Upload:** The user uploads a video through the frontend interface, initiating the processing pipeline.
- 2) **Temporary Storage:** The uploaded video is stored on the backend server to enable controlled and efficient processing.
- 3) **Audio Extraction:** The video is processed to extract the audio stream using a multimedia processing framework, ensuring compatibility with subsequent speech analysis.
- 4) **Transcription:** The extracted audio is converted into text using deep learning-based speech recognition techniques, generating a transcript representation of the video content [28].
- 5) **Validation:** The generated transcript is evaluated using quality metrics to ensure that it contains sufficient semantic information for further processing.
- 6) **SEO Generation:** Natural language processing and information extraction techniques are applied to the validated transcript to generate structured SEO outputs such as summaries, keywords, and metadata [1], [2].
- 7) **Trend Mapping:** A content-driven scoring mechanism evaluates semantic features such as keyword relevance and lexical diversity to estimate engagement potential, inspired by existing tagging and recommendation approaches [5], [31].
- 8) **Visualization:** The final outputs are presented through an interactive dashboard, enabling users to interpret and utilize the generated SEO insights effectively.

VI. IMPLEMENTATION

A. Technology Stack

The system is implemented using a combination of modern web technologies and AI-based processing tools:

- **Frontend:** React.js for building an interactive user interface with asynchronous API communication
- **Backend:** Flask (Python) for handling server-side logic and REST API integration

- **Audio Processing:** FFmpeg for extracting audio streams from video files
- **Speech Recognition:** Whisper model for speech-to-text transcription
- **Generative AI:** Gemini API for generating SEO-related outputs
- **Deployment:** Local server-based deployment with RESTful communication

B. Backend Processing Pipeline

The backend implements a sequential processing pipeline that transforms video input into structured outputs. The major stages include:

- Video ingestion and temporary storage
- Audio extraction using FFmpeg
- Speech-to-text conversion using Whisper
- Transcript validation to ensure quality
- SEO content generation using Gemini API
- Trend analysis for engagement estimation

The system incorporates error handling mechanisms to manage edge cases such as invalid inputs, missing audio, and processing failures, ensuring robustness and reliability.

C. Prompt Engineering

The SEO generation module utilizes structured prompts to guide the generative AI model in producing consistent outputs. The prompt enforces predefined output fields such as summary, keywords, hashtags, and title suggestions.

Information extraction and keyword identification techniques are applied to ensure relevance and coherence in generated outputs [1], [19]. This structured approach reduces variability in output and improves compatibility with downstream processing.

D. System Integration

The frontend and backend communicate via RESTful APIs. The frontend sends video data as multipart requests, while the backend processes the input and returns structured JSON responses. The results are dynamically rendered in the user interface, enabling real-time interaction and visualization.

The modular integration ensures scalability and allows future extensions such as additional analysis modules or external API integration.

VII. EXPERIMENTAL EVALUATION

A. Dataset

The system was evaluated on a curated dataset of 30 videos spanning three categories: educational lectures (10 videos), technical tutorials (10 videos), and general content (10 videos). Video durations ranged from 2 to 15 minutes.

To evaluate system robustness, an additional subset of 5 videos with artificially degraded audio (signal-to-noise ratio < 10 dB) was included. This setup allows assessment of system performance under both ideal and adverse acoustic conditions.

B. Evaluation Metrics

System performance was assessed using the following metrics:

- **Transcript WER:** Word Error Rate computed against manually verified reference transcripts
- **Keyword Precision (KP):** Proportion of generated key-words deemed relevant by evaluators
- **Summary ROUGE-1:** Unigram overlap between generated and reference summaries
- **Processing Latency:** End-to-end time from video upload to result generation

These metrics collectively evaluate transcription quality, semantic relevance, and system efficiency.

C. Results

TABLE I
SYSTEM PERFORMANCE ACROSS VIDEO CATEGORIES

Category	WER (%)	KP (%)	ROUGE-1	Latency (s)
Educational	8.2	84.3	0.61	18.4
Technical	9.7	81.6	0.58	21.2
General Content	12.4	76.8	0.53	19.8
Noisy Audio	34.1	51.2	0.31	23.6

The results presented in Table I demonstrate that VideoIQ achieves strong performance across structured content categories. Keyword precision consistently exceeds 80% for educational and technical videos, indicating effective extraction of relevant semantic features. Similarly, ROUGE-1 scores above 0.55 reflect meaningful summarization quality.

Performance degradation is observed in the presence of noisy audio, where transcription accuracy decreases significantly. This behavior is consistent with known limitations of speech recognition systems under low signal-to-noise conditions [28].

Processing latency increases with video duration and is primarily influenced by the transcription and content analysis stages. Despite this, the system maintains practical response times suitable for real-world usage scenarios.

VIII. DESIGN CONSIDERATIONS

The design of the VideoIQ system is guided by a set of key principles to ensure efficiency, reliability, and future adaptability. These considerations play a crucial role in maintaining system performance and supporting continuous enhancements.

- **Modularity:** The system is structured into independent processing modules, allowing each stage of the pipeline to function and evolve without affecting others.
- **Scalability:** The architecture is designed to efficiently handle videos of varying lengths and increasing workloads without significant performance degradation.
- **Robustness:** Comprehensive error handling and validation mechanisms are implemented to ensure system stability and fault tolerance during execution.
- **Extensibility:** The system supports seamless integration of new features and modules, enabling easy upgrades and incorporation of advanced functionalities.

IX. RESULTS AND DISCUSSION

The system was evaluated across multiple video categories, including educational, technical, and general content, as summarized in Table I.

For videos with clear and structured audio, the system demonstrated strong performance in generating relevant summaries and keywords. Keyword precision exceeded 80% for educational and technical categories, indicating effective extraction of semantically meaningful information. Similarly, ROUGE-1 scores above 0.55 suggest that the generated summaries retain substantial overlap with reference summaries.

The trend analysis module produced consistent engagement estimates that aligned with content richness and lexical diversity. Videos with higher vocabulary variation and well-defined topics yielded higher trend scores, supporting the effectiveness of the proposed feature-based scoring mechanism.

However, performance degradation was observed in videos with low audio quality. Increased background noise and reduced speech clarity negatively impacted transcription accuracy, which in turn affected downstream SEO generation. This behavior is consistent with known challenges in speech processing under noisy conditions [28].

Processing latency ranged between 15–25 seconds for medium-length videos and was primarily influenced by the transcription and content analysis stages. Despite this, the system maintains practical response times suitable for real-world applications.

Overall, the results indicate that VideoIQ effectively automates SEO generation for video content, with performance largely dependent on input audio quality and content clarity.

X. CONCLUSION

This paper presented VideoIQ, a multimodal video SEO intelligence system designed to automate the generation of search-optimized metadata directly from video content. The system integrates speech-based transcription, natural language processing, and content-driven trend analysis within a unified end-to-end pipeline.

Experimental evaluation demonstrated strong performance for videos with clear audio across educational and technical domains, with keyword precision exceeding 80% and ROUGE-1 scores above 0.55. These results validate the effectiveness of the proposed approach in generating meaningful and actionable SEO insights.

The system reduces manual effort in content optimization and provides a scalable solution for improving video discoverability. However, its performance is influenced by audio quality and reliance on external processing modules.

Future work will focus on three key directions: (i) integration of real-time trend data to enhance the trend scoring model, (ii) incorporation of visual content analysis to support non-speech videos, and (iii) extension to multilingual processing for broader applicability.

Overall, VideoIQ demonstrates the potential of combining multimodal analysis and AI-driven techniques for intelligent content optimization in modern video platforms.

ACKNOWLEDGMENT

The authors would like to thank the Department of Computer Science and Engineering, Keshav Memorial Institute of Technology, for providing the necessary resources and support for this research.

REFERENCES

- [1] M. H. A. Abdullah et al., "Systematic Literature Review of Information Extraction From Textual Data," *IEEE Access*, vol. 11, pp. 10562, 2023.
- [2] S. A. Alshalif et al., "Alternative Relative Discrimination Criterion Feature Ranking Technique," *IEEE Access*, vol. 11, pp. 71755, 2023.
- [3] A. Arca et al., "Automated Tag Enrichment by Semantically Related Trends," 2023.
- [4] S. Bostan et al., "Improving Ranking Using Hybrid Custom Embedding Models," *Journal of Web Engineering*, 2023.
- [5] E. B. Boztepe et al., "Audio-Visual Content Understanding using Multimodal Deep Learning," *SAUCIS*, vol. 5, no. 2, 2022.
- [6] D. Cao et al., "Hashtag Recommendation for Micro-Videos," *Knowledge-Based Systems*, vol. 203, 2020.
- [7] Q. Cao and H. Huang, "Video Visual Relation Detection," *IEEE TKDE*, vol. 35, no. 12, 2023.
- [8] X. Chen et al., "Graph Neural Networks for Tag Recommendation," 2023.
- [9] Q. Cheng et al., "Cross-Modal Retrieval Framework," *IEEE TCSVT*, vol. 34, no. 7, 2024.
- [10] X. Du et al., "Personalized Video Recommendation," *IEEE TKDE*, vol. 32, no. 3, 2020.
- [11] F. S. El-metwally et al., "Audio Tagging Framework," *Journal of Advances in IT*, vol. 14, no. 3, 2023.
- [12] Z. Fei et al., "Pairwise Metric Learning for Tag Recommendation," *IEEE Access*, vol. 11, 2023.
- [13] M. Han et al., "Video Object Detection," *IEEE TIP*, vol. 33, 2024.
- [14] M. Hemashree et al., "Semantic AI for Multimedia Tagging," *IEEE Access*, vol. 13, 2025.
- [15] S. Ilyas and H. U. Rehman, "Deep Learning for Video Tagging," 2023.
- [16] W. Jianfeng and S. Xiaowen, "Video Recommendation using Multi-Attention," *ICCEA*, 2023.
- [17] G.-W. Kim et al., "Metadata Extraction using Keyword Mapping," *IEEE Access*, 2024.
- [18] M. K. Najafabadi et al., "Tag Recommendation using Word Embedding," *IEEE SAMI*, 2021.
- [19] M. Nadim et al., "Keyword Extraction Tools Comparison," *IEEE Access*, 2023.
- [20] H. Park et al., "Audio Metadata Tagging using Deep Learning," 2023.
- [21] H. Park et al., "Automatic Movie Tag Generation," *Applied Sciences*, 2022.
- [22] A. A. Patwardhan et al., "ViTag: Automatic Video Tagging," *IEEE BigMM*, 2019.
- [23] H. Ranganathan et al., "Multimodal Emotion Recognition," 2023.
- [24] E. Santos-Neto et al., "Boosting Video Popularity via Tag Selection," 2023.
- [25] F. Schmid et al., "Audio Models using CNN," *IEEE/ACM TASLP*, 2024.
- [26] H. Shrimali and R. Saxena, "Content-Based Video Recommendation," *ICCT*, 2023.
- [27] S. Sun et al., "Answer Ranking using Pre-Trained Models," *IEEE Access*, 2025.
- [28] P. Vitolo et al., "Audio Feature Extraction," *IEEE SPL*, 2024.
- [29] C. Yang et al., "Multimodal Hashtag Recommendation," *IEEE Access*, 2020.
- [30] H. Ye et al., "Extreme Multi-Label Text Classification," *IEEE TKDE*, 2024.
- [31] R. Zhou et al., "Video Tag Recommendation for Mobile Systems," *IEEE Access*, 2020.