॥वसुधैव कुटुम्बकम्॥

SYMBIOSIS INSTITUTE OF TECHNOLOGY, PUNE

Pune – 412115, Maharashtra State, India

https://www.sitpune.edu.in/

DEPARTMENT OF ARTIFICIAL INTELLIGENCE & MACHINE LEARNING

A PROJECT REPORT

ON

"Web-Video Summarization using Titles"

A project report submitted in partial fulfillment of the requirements for the degree of Bachelor of Technology in Artificial Intelligence & Machine Learning

BACHELOR OF TECHNOLOGY IN ARTIFICIAL INTELLIGENCE & MACHINE LEARNING

Submitted By

VAISHNAVI PATIL-21070126108

VIHAN CHORADA-21070126112

UTSAV KARLA-21070126105

SOHOM JANA-21070126092

UNDER THE GUIDANCE OF

Prof. Nivedita Mishra

Prof. Gargi Joshi

Designation

SYMBIOSIS INSTITUTE OF TECHNOLOGY, PUNE

Pune – 412115, Maharashtra State, India
https://www.sitpune.edu.in/

DEPARTMENT OF ARTIFICIAL INTELLIGENCE & MACHINE LEARNING

CERTIFICATE

This is to certify that the Project work entitled "**Title of the Project**" is carried out by the **Name of the Student,** in partial fulfillment for the award of the degree of **Bachelor of Technology** in **Artificial Intelligence & Machine Learning**, Symbiosis International (Deemed University), Pune during the academic year 2024-2025.

Name and Signature of the Guide	Name and Signature of the Co-Guide	Dr. Shruti Patil Head, Department of AI&ML

DECLARATION

I hereby declare that the project titled "**Web-Video Summarization using Titles**" submitted to Symbiosis Institute of Technology, Constituent of Symbiosis International (Deemed University) Pune for the award of the degree of Bachelor of Technology in Artificial Intelligence & Machine Learning is a result of original research carried out by me. I understand that my report may be made electronically available to the public. It is further declared that the project report or any part thereof has not been previously submitted to any University or Institute for the award of any degree or diploma.

Name(s) of Student(s): Vihan Chorada, Vaishnavi Patil, Utsav Karla, Sohom Jana

PRN: 21070126112,21070126108,21070126105,21070126092

Degree: Bachelor of Technology in AI&ML

Department: Artificial Intelligence & Machine Learning

Title of the project : **Web-Video Summarization using Titles**

_____

(Signatures of the Students)

Date: 11/11 /2024

ACKNOWLEDGEMENT

I want to express my sincere gratitude to everyone who supported me throughout this project. First and foremost, I would like to thank my project guide, **Dr.Nivedita Mishra and Dr.Gargi Joshi**, for her valuable guidance, encouragement, and feedback. She has been a constant source of inspiration and motivation for me.

I would also like to thank the head of the department, **Dr. Shruti Patil**, for providing me with the necessary facilities and resources for conducting this project. I am grateful to her for providing constant support and advice.

I want to acknowledge the contribution of my project team members, **Vihan Chorada, Utsav Karla, Sohom Jana, Vaishnavi Patil**, who have worked hard and cooperated with me in every project stage. They have been accommodating and supportive throughout this journey.

I would also like to thank my family and friends for their love, care, and support. They have always been there for me in times of need and stress. They have encouraged me to pursue my passion and achieve my goals.

Lastly, I thank **Symbiosis Institute of Technology Pune** for allowing me to work on this project and enhance my skills and knowledge. I am proud to be a part of this prestigious institution.

TABLE OF CONTENTS

Introduction

With the exponential growth in web video content, managing and understanding this information efficiently has become essential. Video summarization aims to create concise yet meaningful summaries, making it easier to grasp content quickly. By using a multimodal approach—incorporating video frames, titles, annotations, and thumbnails—this project explores enhanced techniques for video summarization to capture diverse video attributes for better summarization. Video summarization is a challenging problem in part because knowing which part of a video is important requires prior knowledge about its main topic. We present TVSum, an unsupervised video summarization framework that uses title-based image search results to find visually important shots. The sheer amount of video available online has increased the demand for efficient ways to search and retrieve desired content. Currently, users choose to watch a video based on various metadata, e.g., thumbnail, title, description, video length, etc. This does not, however, provide a concrete sense of the actual video content, making it difficult to find desired content quickly. Video summarization aims to provide this information by generating the gist of a video, benefiting both the users and companies that provide video streaming and search.

LITERATURE REVIEW

SUMMARY	AUTHOR	YEAR	KEY FINDINGS	METHEDOLOGY	LIMITATION
We present TVSum, an unsupervised video summarization framework that uses title-based image search results to find visually important shots.	Yale Song, Jordi Vallmitjana, Amanda Stent, A. Jaimes	2015	we developed a novel co-archetypal analysis technique that learns canonical visual concepts shared between video and images, but not in either alone, by finding a joint-factorial representation of two data sets.	Unsupervised video summarization framework called TVSum - Leveraging title-based image search results to identify visually important video shots - Developing a novel co-archetypal analysis technique to address noise and variance in the title-based image search results	he co-archetypal analysis technique was developed to address the challenge of noise and variance in the title-based image search results, suggesting this may be a limitation of the approach - The small size of the TVSum50 dataset (50 videos) may be a limitation, as the authors do not mention whether the approach has been tested on a larger dataset
The paper presents a framework for summarizing web videos based on key shots to provide an	Richang Hong, Jinhui Tang, Hung-Khoon Tan, Shuicheng Yan, C. Ngo, Tat-Seng Chu	2009	his paper presents a novel solution by mining and threading "key" shots, which can provide an overview of	Identify "key shots" by detecting near-duplicate keyframes, ranking them based on informativeness,	The framework only provides a static summary, and dynamic video skimming may be a limitation that

overview of the main contents.			main contents of videos at a glance, by summarizing a large set of diverse videos.	and arranging them in chronological order. 3. Formulate the summarization as an optimization problem that balances the relevance of the key shots and a user-defined skimming ratio.	could be improved upon.
A multimodal approach using closed captions and speech signals to summarize and index news video.	Sooyoung Kim Shin, Kwangjae Lim, Kwonhue Choi, K. Kang	2002	The proposed method exploits the closed caption data to locate semantically meaningful highlights in a news video and speech signals in an audio stream to align the closed caption data with the video in a time-line.	Aligning the closed captions with the video timeline - Describing the highlights using the MPEG-7 Summarization	The method heavily relies on the availability of accurate closed caption data. In cases where captions are missing, incomplete, or inaccurate, the method's effectiveness is compromised.
Multi-modal summarization of key events and top players in sports tournament videos	D. Tjondronegoro, Xiaohui Tao, Johannes Sasongko, C. H. Lau	2011	This paper aims to address this limitation using a novel multimodal summarization framework that is based on sentiment	Sentiment analysis and player popularity analysis to automatically annotate and visualize the key events and key	Previous work has only used time-stamped web match reports synchronized with video, but web and social media articles

			analysis and players' popularity. It uses audiovisual contents, web articles, blogs, and commentators' speech to automatically annotate and visualize the key events and key players in a sports tournament coverage	players in the sports tournament coverage	without timestamps have not been fully leveraged
MMSS provides a domain-independent, graph-based framework for multi-modal story-oriented video summarization.	Jia-Yu Pan, Hyung-Jeong Yang, C. Faloutsos	2004	We propose multi-modal story-oriented video summarization (MMSS) which, unlike previous works that use fine-tuned, domain-specific heuristics, provides a domain-independent, graph-based framework	Multimodal data integration to uncover correlations between different information modalities - Domain-independent approach, not relying on fine-tuned, domain-specific heuristics	Graph-based modeling and multi-modal data correlation require extensive resources, affecting scalability.
Read, Watch, Listen, and Summarize: Multi-Modal	Haoran Li, Junnan Zhu, Cong Ma, Jiajun Zhang,	2019	The key idea is to bridge the semantic gaps between multi-	Learning joint representations of text and images using	The system was primarily trained on English and

Summarization for Asynchronous Text, Image, Audio and Video	Chengqing Zong		modal content. Audio and visual are main modalities in the video.	neural networks, and then using text-image matching or multi-modal topic modeling to ensure the generated summary covers important visual information	Chinese news datasets, limiting its adaptability to other languages or content domains without further training.
A multimodal abstractive summarization model integrates information from video and audio transcripts to generate coherent textual summaries for open-domain videos.	Shruti Palaskar, Jindřich Libovický, Spandana Gella, Florian Metze	2019	nlike the traditional text news summarization, the goal is less to “compress” text information but rather to provide a fluent textual summary of information that has been collected and fused from different source modalities, in our case video and audio transcripts (or text).	Experiments on the How2 corpus of instructional videos - Proposal of a new evaluation metric called "Content F1" that measures semantic adequacy of the summaries	he study presents a new model and conducts pilot experiments, suggesting that further research and validation is needed

Problem Statement, Goal & Motivation

□ Problem Statement: In the current digital landscape, video content is vast and growing, which makes it challenging for users to consume it efficiently. There is a need for summarization techniques that can condense video content into key moments, enhancing viewer accessibility without sacrificing the essence of the content.

□ Goal: The goal is to develop a system that can generate concise and accurate video summaries by leveraging various input modalities such as video frames, textual information, and user-generated annotations.

□ Motivation: This approach addresses the limitations of unimodal summarization techniques by adopting a multimodal strategy, which combines diverse data sources. Multimodal methods are motivated by the potential to capture a richer understanding of video content, thereby improving summarization quality and ensuring better user engagement. By bridging information from textual, visual, and annotated data, the project aims to set new standards for creating reliable, context-aware summaries.

Objectives

To identify and extract significant features from each modality—visual content, textual titles, and user annotations—that contribute to meaningful summarization.

To apply and test various data fusion techniques, including early, late, and hybrid fusion, to determine which method provides the most coherent summaries.

To evaluate the effectiveness of multimodal fusion techniques against unimodal approaches in terms of relevance, accuracy, and user satisfaction.

To optimize processing to ensure the approach is computationally efficient and can scale across various video types and genres.

To examine how multimodality aids in refining and enhancing the summarization process for applications in fields like content recommendation, media editing, and content management.

Dataset Details

Dataset Overview: The dataset used in this project is the TVSum dataset, which consists of 50 videos across diverse categories like news, sports, and vlogs, each paired with user-generated titles and relevance annotations.

Annotations: Each video is annotated with user importance scores for different segments, which serve as a basis for training and evaluating the summarization model.

Additional Modalities: The dataset includes accompanying thumbnails that visually represent key scenes. Titles are stored in ydata-tvsum50-info.tsv, while annotations are in ydata-tvsum50-anno.tsv, providing user preferences and insights on video relevance.

Preprocessing Requirements: Preprocessing involves reading titles, relevance annotations, and extracting or resizing thumbnails. Video frame extraction and feature encoding are additional steps necessary for effective multimodal fusion.

Modalities Chosen for the Task

• Textual Information: The titles provide an overarching context, capturing user intent and summarizing the central theme of each video.

• Visual Information: Thumbnail images encapsulate important visual cues from the video and are essential in reinforcing the visual appeal and relevance of summaries.

• Annotation Data: User relevance scores for each segment help gauge important moments, allowing for user-focused summarization that aligns with viewer expectations.

• Multimodal Fusion Advantage: Using all three modalities enables the summarization system to capture nuances and synthesize information more effectively than unimodal approaches.

Processing of Individual Modalities

Text Processing: Titles are preprocessed using tokenization and transformed into numerical vectors using techniques such as TF-IDF for meaningful representation.

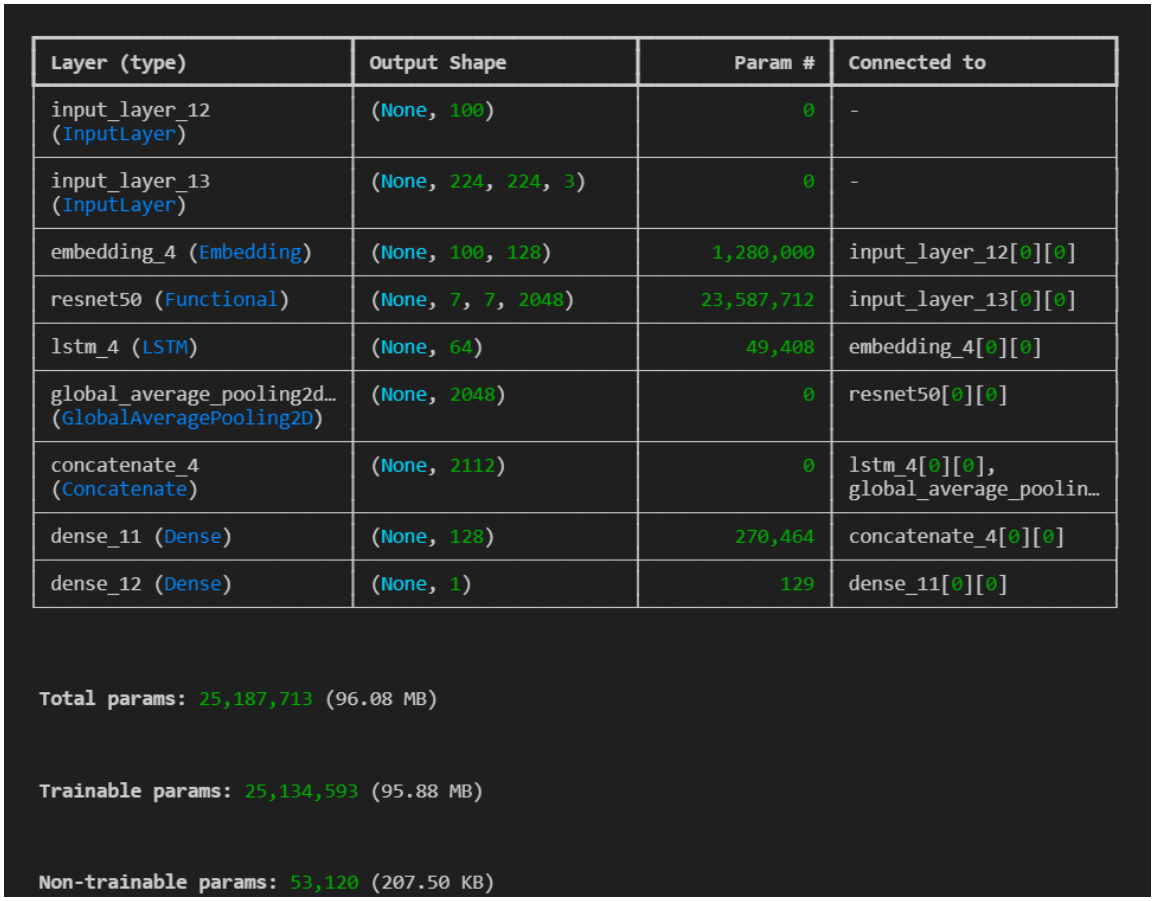
Visual Processing: Thumbnail images are resized and feature extraction is performed to reduce dimensionality while retaining significant visual patterns, using models like ResNet or EfficientNet.

Annotation Processing: The relevance scores are normalized, categorized, and aligned with the video segments, making them compatible with other modalities during fusion.

Challenges and Preprocessing Requirements: Each modality has distinct processing needs; balancing these is crucial for the smooth integration of modalities in fusion processes.

Multimodal Data Fusion

- **Early Fusion**: Combines features from text, visual, and annotation modalities at the initial stages. By merging raw or early-processed features, early fusion allows the system to analyze content holistically but may become computationally intensive.



Layer (type)	Output Shape	Param #	Connected to
input_layer_12 (InputLayer)	(None, 100)	0	-
input_layer_13 (InputLayer)	(None, 224, 224, 3)	0	-
embedding_4 (Embedding)	(None, 100, 128)	1,280,000	input_layer_12[0][0]
resnet50 (Functional)	(None, 7, 7, 2048)	23,587,712	input_layer_13[0][0]
lstm_4 (LSTM)	(None, 64)	49,408	embedding_4[0][0]
global_average_pooling2d… (GlobalAveragePooling2D)	(None, 2048)	0	resnet50[0][0]
concatenate_4 (Concatenate)	(None, 2112)	0	lstm_4[0][0], global_average_poolin…
dense_11 (Dense)	(None, 128)	270,464	concatenate_4[0][0]
dense_12 (Dense)	(None, 1)	129	dense_11[0][0]

Total params: 25,187,713 (96.08 MB)

Trainable params: 25,134,593 (95.88 MB)

Non-trainable params: 53,120 (207.50 KB)

- **Late Fusion**: Involves processing each modality separately to create individual summaries, then merging them at the end. This approach is computationally simpler and allows flexibility in assigning importance to each modality based on its contribution.

Layer (type)	Output Shape	Param #	Connected to
input_layer_3 (InputLayer)	(None, 100)	0	-
input_layer_4 (InputLayer)	(None, 224, 224, 3)	0	-
embedding_1 (Embedding)	(None, 100, 128)	1,280,000	input_layer_3[0][0]
resnet50 (Functional)	(None, 7, 7, 2048)	23,587,712	input_layer_4[0][0]
lstm_1 (LSTM)	(None, 64)	49,408	embedding_1[0][0]
global_average_pooling2d… (GlobalAveragePooling2D)	(None, 2048)	0	resnet50[0][0]
dense_2 (Dense)	(None, 1)	65	lstm_1[0][0]
dense_3 (Dense)	(None, 1)	2,049	global_average_poolin…
average (Average)	(None, 1)	0	dense_2[0][0], dense_3[0][0]

Total params: 24,919,234 (95.06 MB)

Trainable params: 24,866,114 (94.86 MB)

Non-trainable params: 53,120 (207.50 KB)

- **Hybrid Fusion**: Combines the benefits of both early and late fusion by merging modalities at both initial and final stages, creating a balance between efficiency and comprehensiveness.

Layer (type)	Output Shape	Param #	Connected to
input_layer_6 (InputLayer)	(None, 100)	0	-
input_layer_7 (InputLayer)	(None, 224, 224, 3)	0	-
embedding_2 (Embedding)	(None, 100, 128)	1,280,000	input_layer_6[0][0]
resnet50 (Functional)	(None, 7, 7, 2048)	23,587,712	input_layer_7[0][0]
lstm_2 (LSTM)	(None, 64)	49,408	embedding_2[0][0]
global_average_pooling2d… (GlobalAveragePooling2D)	(None, 2048)	0	resnet50[0][0]
concatenate_1 (Concatenate)	(None, 2112)	0	lstm_2[0][0], global_average_poolin…
dense_4 (Dense)	(None, 128)	270,464	concatenate_1[0][0]
dense_5 (Dense)	(None, 64)	4,160	lstm_2[0][0]
dense_6 (Dense)	(None, 64)	131,136	global_average_poolin…
concatenate_2 (Concatenate)	(None, 256)	0	dense_4[0][0], dense_5[0][0], dense_6[0][0]
dense_7 (Dense)	(None, 128)	32,896	concatenate_2[0][0]
dense_8 (Dense)	(None, 1)	129	dense_7[0][0]

Total params: 25,355,905 (96.73 MB)

- **Application and Evaluation**: Each fusion type is tested to determine how well it performs with the multimodal data and to assess the quality of summaries generated.

Comparison of Different Fusion Techniques with Individual Modalities

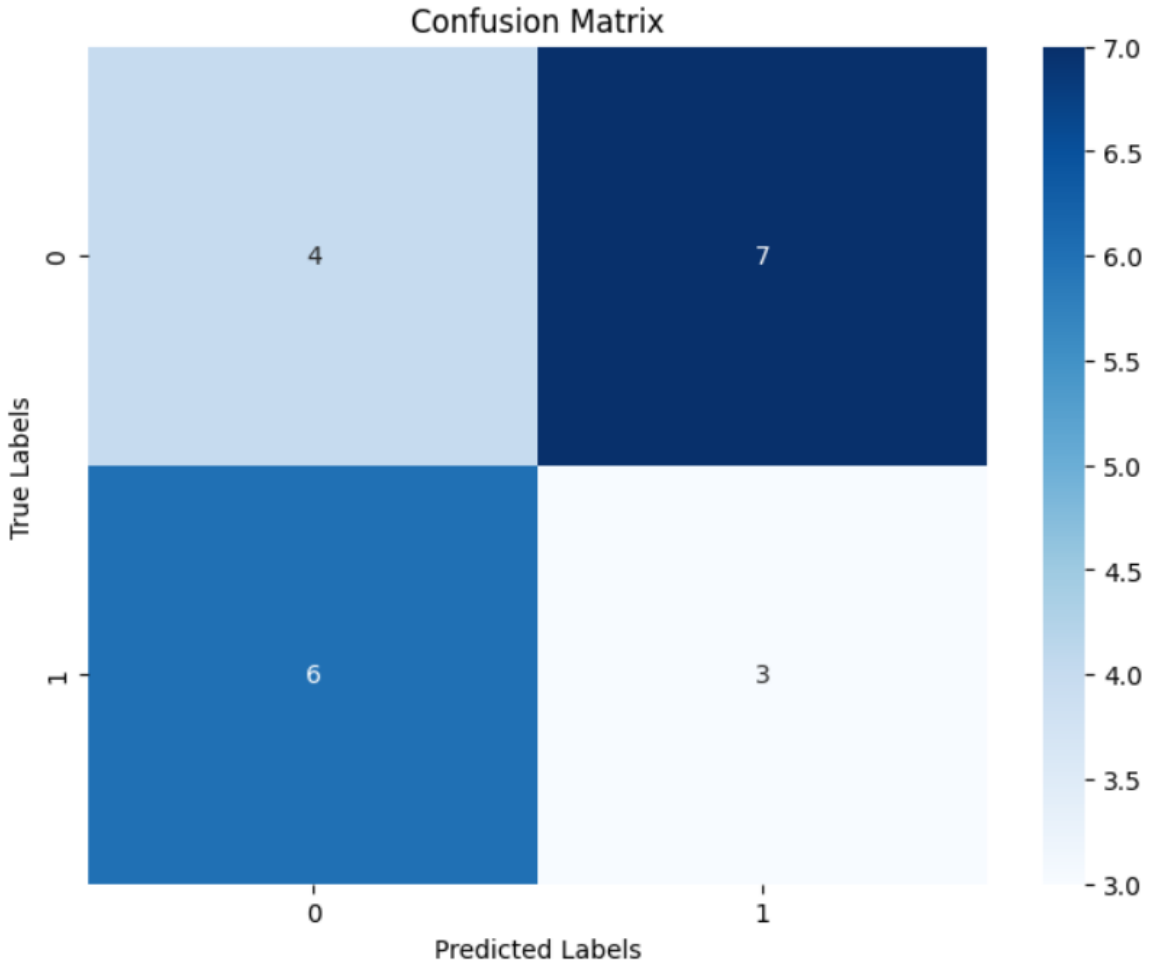
 Unimodal Baseline: Individual modality performance is assessed to serve as a baseline. Text-only, visual-only, and annotation-only summarizations are created and evaluated on coherence and relevance.

 Fusion Comparison: Early, late, and hybrid fusion results are compared against unimodal summaries, with metrics including summary relevance, viewer satisfaction, and alignment with annotated scores.

 Findings: The fusion techniques are more effective than individual modalities in capturing diverse aspects of video content. This comparison highlights that while each modality provides unique information, fusing them leads to a more comprehensive and meaningful summary.

Fusion Evaluation Results Comparing with Individual Modalities

- **Evaluation Metrics**: Metrics like precision, recall, F1-score, and user satisfaction ratings are used to assess summary quality.
- **Results Summary**: Results indicate that hybrid fusion provides the best balance between computational efficiency and summary relevance. Late fusion also performs well but sometimes lacks the detail captured by early fusion.
- **Insights on Performance**: The multimodal fusion approach captures a broader range of information, making the summaries more robust, contextually accurate, and reflective of user preferences.



Analysis of Observed Results

 Best Fusion Technique: Hybrid fusion is identified as the most effective technique for this task, as it combines the early-stage comprehensiveness with the flexibility of late fusion.

 Multimodality Benefits: Integrating multiple modalities leads to significant improvements in summary quality, making it more aligned with user interests. The combination of modalities enhances the depth and context of summaries, which is not achievable through individual modalities.

 Challenges Addressed: The results confirm that multimodal fusion addresses common challenges in summarization, such as balancing content brevity with detail and aligning with diverse user expectations.

Comparison of Early Fusion, Late Fusion, Hybrid Fusion Accuracy

```
Early Fusion:
1/1 ━━━━━━━━━━━━━━━━━━━━ 3s 3s/step - accuracy: 0.3500 - loss: 0.6940
Late Fusion:
1/1 ━━━━━━━━━━━━━━━━━━━━ 3s 3s/step - accuracy: 0.4500 - loss: 0.7050
Hybrid Fusion:
1/1 ━━━━━━━━━━━━━━━━━━━━ 4s 4s/step - accuracy: 0.5500 - loss: 0.7727

[0.7726967930793762, 0.550000011920929]
```

Conclusion

Summary: This project successfully demonstrates the effectiveness of multimodal fusion techniques for video summarization. By integrating textual, visual, and annotation data, the project achieves higher-quality summaries that meet user preferences and provide a coherent overview of content.

Future Directions: Future research may explore additional modalities, such as audio or emotion detection, to further refine video summarization. Additionally, the use of real-time processing models could enhance scalability and applicability across larger datasets.

Broader Implications: The findings suggest that multimodal approaches could greatly benefit fields requiring video summarization, from media production to educational platforms, making content more accessible and engaging for users.