Hybrid GNN-Driven Shot Selection for Multimodal Video Summarization

L Moryakantha*
*Department of AIML
R.V. College of Engineering
Bangalore, India
lmoryakantha.ai24@rvce.edu.in

Mahantesh P B*
*Department of AIML
R.V. College of Engineering
Bangalore, India
mahanteshpb.ai24@rvce.edu.in

Dr. K. Viswavardhan Reddy†
†Department of AIML
R.V. College of Engineering
Bangalore, India
viswavardhank@rvce.edu.in

***Abstract*—The rapid growth of long-form video content across digital platforms has created a strong demand for automated video summarization techniques that are both efficient and semantically meaningful. Existing approaches often rely on frame-level or linear sequence-based models, which suffer from redundancy, scalability limitations, and a reduced ability to capture long-range contextual relationships. To address these challenges, this paper proposes a multimodal graph-based video summarization framework operating at the shot level.**

Visual features are extracted using a Vision Transformer (ViT), while audio features are obtained through the self-supervised HuBERT model to capture speech-related semantics. The extracted multimodal features are represented as nodes in a graph, where inter-shot dependencies and contextual relationships are modeled using a Graph Attention Network (GATv2). This graph-based formulation enables effective relational reasoning beyond sequential constraints. To further enhance summary quality and efficiency, only the most important shots are selectively processed using speech-to-text conversion followed by abstractive text summarization.

Experimental results demonstrate stable convergence, effective importance ranking, reduced redundancy, and improved summary relevance when compared with traditional sequence-based video summarization approaches, highlighting the effectiveness of the proposed framework for long-form multimedia content.

***Index Terms*—Video Summarization, Graph Neural Networks, Multimodal Learning, Vision Transformer, Graph Attention Network**

I. Introduction

The exponential growth of video data across digital platforms such as online learning portals, media archives, entertainment platforms, and social networks has created a strong demand for efficient video understanding and content management solutions. Long-form videos are increasingly used for education, surveillance, corporate communication, and media streaming, making manual browsing and content extraction both time-consuming and impractical. Automatic video summarization aims to address this challenge by generating concise summaries that preserve the most important semantic content of videos while significantly reducing viewing time.

Traditional video summarization techniques primarily focused on extracting key frames or short clips based on low-level visual features such as color, motion, and texture. While computationally efficient, these methods lacked semantic awareness and often resulted in redundant or visually incoherent summaries. As video lengths and content complexity increased, such heuristic-based approaches became insufficient for capturing high-level contextual information.

Recent advances in deep learning have significantly improved video summarization performance by enabling the extraction of higher-level representations. Convolutional neural networks and recurrent neural networks have been widely used to model spatial and temporal information in videos. However, these models typically rely on sequential processing and struggle to capture long-range dependencies in long-form video content. Transformer-based architectures have addressed some of these limitations through self-attention mechanisms, enabling global context modeling. Nevertheless, directly applying Transformers to long video sequences is computationally expensive and often ignores explicit relational structures between video segments.

Moreover, most existing approaches predominantly rely on visual information, overlooking the rich semantic cues present in audio and speech modalities. Incorporating multimodal information, such as speech emphasis and audio context, has been shown to improve summarization quality, particularly for lecture-style and presentation-based videos. However, effective integration of multimodal features with contextual reasoning remains a challenging research problem.

To overcome these limitations, this work proposes a shot-level multimodal video summarization framework that explicitly models inter-shot relationships using graph-based learning. By representing video shots as nodes in a graph and leveraging graph attention mechanisms for importance estimation, the proposed approach enables contextual reasoning across both local and long-range video segments. This design improves summary coherence, reduces redundancy, and enhances scalability for long-form video summarization tasks.

II. Background Study

Early research in video summarization primarily focused on key-frame extraction and shot boundary detection using handcrafted features such as color histograms, motion vectors, and edge descriptors. While these methods were computationally efficient, they lacked semantic understanding and often produced redundant or visually incoherent summaries.

To overcome these limitations, machine learning-based approaches were introduced, incorporating clustering, ranking, and optimization techniques to improve summary quality.

With the advancement of deep learning, convolutional neural networks (CNNs) and recurrent neural networks (RNNs) became widely adopted for video summarization tasks. CNN-based models enabled higher-level visual feature extraction, while RNNs and Long Short-Term Memory (LSTM) networks modeled temporal dependencies across video frames. Although these methods demonstrated improved performance, they relied heavily on sequential processing and were limited in capturing long-range dependencies in lengthy videos.

Transformer-based architectures further advanced the field by introducing self-attention mechanisms capable of modeling global context. Vision Transformers (ViT) enabled effective global visual representation learning, while attention-based sequence models improved temporal reasoning. However, Transformers applied directly to long video sequences are computationally expensive and often ignore the relational structure between video segments.

III. Literature Survey

Video summarization has evolved from traditional heuristic-based methods to advanced deep learning approaches driven by the exponential growth of multimedia data. Early systems primarily relied on low-level handcrafted features such as color histograms, edge density, motion vectors, and clustering techniques to identify key frames. Although these methods were computationally efficient, they lacked semantic understanding and often produced redundant summaries.

With the advent of deep learning, convolutional neural networks (CNNs) and recurrent neural networks (RNNs) were introduced to capture higher-level visual representations and temporal dependencies. However, these models assumed linear temporal relationships and struggled with long-range dependencies in extended videos.

Transformer-based architectures marked a significant breakthrough by enabling global context modeling using self-attention mechanisms. Vision Transformers (ViT) demonstrated strong performance in visual feature extraction, while self-supervised speech models such as HuBERT enabled robust audio feature learning without labeled data. Despite these advancements, many existing approaches failed to explicitly model relationships between video segments.

Graph Neural Networks (GNNs) emerged as a promising solution for relational reasoning. By modeling video segments as nodes and their interactions as edges, GNNs enabled contextual understanding beyond sequential constraints. However, limited research has explored the integration of multimodal Transformers with graph attention mechanisms for video summarization, motivating the proposed work.

A. Survey of Related Work

1) Vaswani et al. (2017) introduced the Transformer architecture based on self-attention, demonstrating superior performance in sequence modeling tasks. This work established the foundation for modern Transformer-based models used in vision and language processing.
2) Dosovitskiy et al. (2021) proposed the Vision Transformer (ViT), which applied Transformer architectures directly to image patches. The study showed that ViT effectively captures global visual context, making it suitable for visual feature extraction in video summarization.
3) Hsu et al. (2021) presented HuBERT, a self-supervised speech representation learning framework. The model successfully learned semantic and phonetic information from raw audio, enabling robust audio feature extraction without manual labeling.
4) Zhang et al. (2016) explored key-frame extraction methods based on visual saliency and clustering. While effective for short videos, the approach struggled with long-form videos due to redundancy.
5) Mahasseni et al. (2017) proposed adversarial learning for unsupervised video summarization. The method reduced labeling requirements but suffered from unstable training and limited semantic understanding.
6) Zhao et al. (2018) used Long Short-Term Memory (LSTM) networks for video summarization. Although temporal dependencies were modeled, long-range context handling remained limited.
7) Zhou et al. (2018) introduced attention-based sequence models for video summarization, improving focus on important frames but still relying on linear temporal assumptions.
8) Li et al. (2019) incorporated audio-visual fusion for summarization, demonstrating improved results compared to visual-only methods.
9) Chu et al. (2020) applied reinforcement learning to optimize summary selection. However, the reward design was complex and computationally expensive.
10) Veličković et al. (2018) introduced Graph Attention Networks (GAT), enabling attention-based aggregation of graph node features. This work motivated graph-based modeling of video shots.
11) Wu et al. (2020) applied Graph Neural Networks for video understanding, highlighting their effectiveness in modeling inter-segment relationships.
12) Lei et al. (2020) proposed hierarchical video summarization using shot-level representations, reducing redundancy in summaries.
13) Radford et al. (2022) introduced Whisper, a large-scale speech recognition model that demonstrated robust performance across noisy and multilingual speech scenarios.
14) Lin et al. (2021) explored multimodal Transformers for video-text understanding, emphasizing the importance of cross-modal learning.
15) Chen et al. (2022) combined visual and audio cues for summarization using deep fusion strategies.
16) Zhang et al. (2022) proposed graph-based summariza-

tion using temporal similarity graphs, improving contextual coherence.

17) Liu et al. (2023) explored long-video summarization using sparse attention mechanisms to reduce computational complexity.
18) OpenAI Documentation (2023) provided implementation details and performance benchmarks for Transformer-based models.
19) PyTorch Documentation (2023) described deep learning frameworks and optimization strategies for large-scale models.
20) FFmpeg Documentation (2023) detailed efficient video decoding and audio-video processing techniques used in multimedia applications.

B. Motivation and Research Gap

From the literature survey, it is evident that existing video summarization methods face several unresolved challenges, including redundancy in summaries, limited contextual reasoning, high computational cost for long videos, and insufficient integration of audio-semantic information. Most approaches either focus on visual features alone or lack an explicit mechanism to model long-range relationships between video shots.

Motivated by these observations, this work proposes a shot-level multimodal video summarization framework that integrates Transformer-based feature extraction with graph-based relational reasoning. By representing video shots as nodes in a graph and modeling temporal and semantic relationships using a Graph Attention Network, the proposed approach aims to generate more coherent, informative, and context-aware summaries.

The proposed framework effectively bridges the gap between multimodal representation learning and relational modeling, making it well-suited for summarizing long-form videos in real-world applications.

IV. Contributions

The main contributions of this paper are summarized as follows:

- A shot-level video representation framework that reduces redundancy and improves temporal coherence compared to frame-level summarization approaches.
- A multimodal feature extraction and fusion strategy that combines Vision Transformer (ViT) based visual representations with HuBERT-based speech-aware audio embeddings.
- A graph-based modeling approach that represents video shots as nodes and explicitly captures temporal and semantic relationships using a Graph Attention Network (GATv2).
- An effective shot importance prediction mechanism that leverages attention-based relational reasoning to identify informative video segments.
- A computationally efficient summarization pipeline that selectively applies speech-to-text conversion and transformer-based abstractive text summarization only to the most relevant shots.

V. Related Work

Early video summarization methods focused on key-frame extraction using handcrafted features such as color histograms and motion vectors. With the advancement of deep learning, CNN-based and RNN-based approaches were introduced to capture higher-level representations and temporal dependencies.

Transformer-based models further improved global context modeling but are computationally expensive for long videos. Multimodal approaches incorporating audio information have shown improved performance; however, they often lack explicit relational modeling. Graph Neural Networks have recently been explored to capture relationships between video segments, motivating the proposed approach.

VI. Problem Definition

Given an input video V, the objective is to generate a concise summary that preserves the most important semantic content while significantly reducing the overall duration of the video. The input video V is first segmented into a sequence of N semantic shots, denoted as $S = \{S_1, S_2, \ldots, S_N\}$, where each shot represents a continuous temporal segment with coherent visual and audio content.

For each shot S_i, multimodal features are extracted from both visual and audio modalities. Visual features capture the spatial and scene-level information, while audio features encode speech and acoustic cues that often indicate semantic importance. In addition to individual shot characteristics, contextual relationships between shots—such as temporal continuity and semantic similarity—play a critical role in determining overall importance.

The primary objective is to learn a function $f(\cdot)$ that maps each shot S_i to a continuous importance score $y_i \in \mathbb{R}$ by jointly considering multimodal features and inter-shot dependencies:

$$y_i = f(S_i, \mathcal{N}(S_i)), \tag{1}$$

where $\mathcal{N}(S_i)$ denotes the neighborhood of shot S_i capturing its contextual relationships within the video.

Based on the predicted importance scores, a subset of shots $\hat{S} \subseteq S$ is selected such that the resulting summary is informative, non-redundant, and temporally coherent under a given summary length constraint. Finally, a coherent textual summary is generated from the selected shots using automatic speech recognition and abstractive text summarization techniques.

VII. Proposed Methodology

This section presents the proposed *multimodal graph-based video summarization framework* in detail. The methodology is designed to operate at the *shot level*, integrate *visual and audio information*, explicitly model *inter-shot relationships*, and generate concise textual summaries using transformer-based language models. An overview of the proposed pipeline is illustrated in Figure 2.

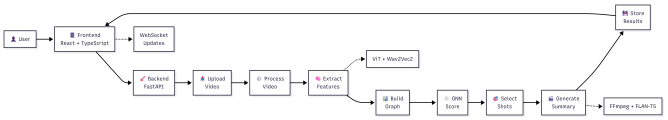

Fig. 1. End-to-end system architecture of the proposed multimodal graph-based video summarization framework, showing frontend–backend interaction, feature extraction, graph-based importance scoring, and summary generation.

A. System Overview

The proposed system follows a modular end-to-end pipeline. Given an input video, the video is first segmented into semantic shots. Each shot is represented using multimodal features extracted from both visual frames and audio signals. These shot representations are modeled as nodes in a graph, where edges encode temporal continuity and semantic similarity. A Graph Attention Network (GATv2) is then employed to predict importance scores for each shot. Finally, the most important shots are selectively transcribed and summarized to generate the final video summary.

B. Shot Detection and Preprocessing

The first step of the proposed framework is to decompose the input video into a sequence of semantically coherent shots. A *shot* is defined as a continuous sequence of frames captured without interruption by a single camera operation. Shot boundary detection is performed using content-based techniques that identify abrupt and gradual visual changes between consecutive frames based on metrics such as color histogram differences, edge variation, and motion discontinuities.

Shot-level segmentation significantly reduces temporal redundancy compared to frame-level processing while preserving semantically meaningful units. This design choice improves computational efficiency and scalability, particularly for long-form videos.

After detecting shot boundaries, each shot undergoes preprocessing to remove redundant frames. Let $\{F_1, F_2, \ldots, F_m\}$ denote the frames within a shot. Frame difference measures and visual similarity scores are used to eliminate near-duplicate frames. From the remaining frames, a single representative frame F_r is selected. For static shots, the middle frame is chosen:

$$F_r = F_{\lfloor m/2 \rfloor}$$

whereas for dynamic shots, the frame with the highest motion energy is selected to ensure maximal visual information.

In parallel, the audio stream corresponding to each shot is extracted using the detected shot start and end timestamps. The extracted audio segments are converted to mono format and resampled to a fixed sampling rate to ensure uniformity across all shots. This temporal alignment of audio and visual data enables effective multimodal feature extraction.

—

C. Visual Feature Extraction

Visual representations are extracted from the representative frame of each shot using a pretrained Vision Transformer (ViT). Unlike convolutional neural networks that primarily capture local spatial patterns, ViT leverages self-attention mechanisms to model global contextual relationships across the entire image.

Each input frame is resized to a fixed resolution and divided into a set of non-overlapping patches. Let an image be split into P patches $\{p_1, p_2, \ldots, p_P\}$. Each patch is flattened and linearly projected into an embedding space. Positional embeddings are added to preserve spatial structure:

$$\mathbf{z}_0 = [\mathbf{x}_{\text{CLS}}; \mathbf{x}_1 + \mathbf{e}_1; \ldots; \mathbf{x}_P + \mathbf{e}_P]$$

The sequence is then processed through multiple Transformer encoder layers:

$$\mathbf{z}_l = \text{TransformerEncoder}(\mathbf{z}_{l-1})$$

The output corresponding to the classification token (`[CLS]`) is used as the compact visual representation of the shot. This vector captures high-level semantic attributes such as objects, scenes, and contextual cues. The resulting visual embedding is a 768-dimensional feature vector:

$$\mathbf{v}_i \in \mathbb{R}^{768}$$

The ViT model is used in a frozen configuration to leverage pretrained knowledge and reduce training complexity.

—

D. Audio Feature Extraction

Audio features are extracted using HuBERT, a self-supervised speech representation learning model capable of encoding both phonetic and semantic speech characteristics. HuBERT learns meaningful representations by predicting masked latent speech units, enabling robust modeling without explicit labeled supervision.

Let A_i denote the audio segment corresponding to the i^{th} shot. This segment is passed through the HuBERT encoder to generate frame-level embeddings:

$$\{\mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_T\}$$

To obtain a fixed-length representation for each shot, mean pooling is applied across the temporal dimension:

$$\mathbf{a}_i = \frac{1}{T} \sum_{t=1}^{T} \mathbf{h}_t$$

The resulting audio feature vector $\mathbf{a}_i \in \mathbb{R}^{768}$ captures speech emphasis, intonation, and semantic content, which are crucial indicators of importance in many video domains such as lectures, meetings, and presentations.

—

E. Multimodal Feature Fusion

Visual and audio modalities provide complementary information; therefore, they are combined to form a unified shot-level representation. Let $\mathbf{v}_i \in \mathbb{R}^{768}$ and $\mathbf{a}_i \in \mathbb{R}^{768}$ denote the visual and audio feature vectors of the i^{th} shot, respectively.

The multimodal feature vector is obtained by concatenation:

$$\mathbf{x}_i = [\mathbf{v}_i \,||\, \mathbf{a}_i] \tag{2}$$

where $||$ denotes vector concatenation. The resulting multimodal embedding $\mathbf{x}_i \in \mathbb{R}^{1536}$ captures both visual semantics and audio context.

This fused representation serves as the node feature in the graph-based modeling stage. By operating on multimodal shot-level embeddings, the framework ensures robust importance estimation even when one modality alone is insufficient, such as visually static but semantically rich speech segments or visually salient scenes with limited audio content.

F. Graph Construction

Each video is modeled as a graph $G = (V, E)$, where each node $v_i \in V$ corresponds to a video shot. Edges are constructed based on the following criteria:

- Temporal adjacency between consecutive shots
- Visual similarity computed using cosine similarity between visual embeddings
- Audio similarity computed using cosine similarity between audio embeddings

This graph structure enables modeling of both local and global relationships between video shots.

G. Shot Importance Prediction Using GATv2

The constructed graph is processed using a Graph Attention Network (GATv2). Each GATv2 layer updates node representations by aggregating information from neighboring nodes using learned attention weights that reflect their relative importance. Multi-head attention is employed to capture diverse relational patterns between shots.

Residual connections, layer normalization, and nonlinear activation functions are applied to stabilize training and prevent over-smoothing. The final node representations are passed through a multilayer perceptron to predict a continuous importance score for each shot.

H. Shot Selection and Text Summarization

Based on the predicted importance scores, shots are ranked, and the top-K most important shots are selected. Audio segments corresponding to these shots are transcribed using Whisper ASR, a robust speech-to-text model.

The resulting transcripts are then summarized using a transformer-based language model such as Flan-T5 or DeepSeek to generate a concise and coherent textual summary of the video. By selectively processing only important shots, the system improves both computational efficiency and summary relevance.

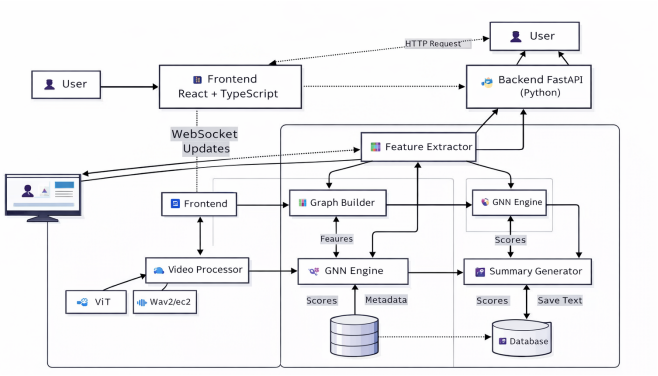

Fig. 2. High-level system architecture of the proposed multimodal graph-based video summarization framework, illustrating frontend–backend interaction, multimodal feature extraction using ViT and HuBERT, graph construction, GAT-based importance scoring, and summary generation.

I. System Architecture Description

Figure 2 illustrates the end-to-end architecture of the proposed AI-based video summarization system. The framework follows a client–server design in which a web-based frontend developed using React and TypeScript interacts with a FastAPI backend that manages video processing and model execution. Real-time progress updates are delivered to the user through WebSocket communication.

Once a video is uploaded, the backend performs preprocessing by decoding the video, segmenting it into semantic shots, and extracting shot-aligned audio segments. Representative frames from each shot are processed using a Vision Transformer to extract global visual features, while corresponding audio segments are analyzed using Wav2Vec2 or HuBERT to obtain speech-aware audio embeddings. These multimodal features are fused and used to construct a graph representation of the video, where nodes correspond to shots and edges encode temporal and semantic relationships.

The constructed graph is processed using a Graph Attention Network to predict importance scores for each shot based on contextual relevance. The most informative shots are then selected, and their audio segments are transcribed using an automatic speech recognition model. The resulting transcripts are summarized using a transformer-based language model such as Flan-T5 to generate a concise textual summary. The final summary is stored by the backend and returned to the frontend for user access.

VIII. Training Strategy

The proposed model is trained to predict continuous importance scores for video shots using a regression-based learning objective. Since shot importance annotations are inherently imbalanced—with a small fraction of shots being highly informative—a weighted Smooth L1 loss function is employed. This loss formulation penalizes large prediction errors while remaining robust to outliers and assigns higher importance to informative shots, thereby improving ranking consistency and summary relevance.

The dataset is divided into training, validation, and test sets using a standard split of 70%, 15%, and 15%, respectively. The training set is used to learn model parameters, while the validation set is used for hyperparameter tuning and early stopping. The test set is reserved exclusively for final performance evaluation to ensure unbiased assessment. This split strategy provides a balanced trade-off between sufficient training data and reliable generalization evaluation.

Optimization is performed using the AdamW optimizer, which decouples weight decay from gradient updates and has been shown to improve convergence and generalization for Transformer- and graph-based architectures. An adaptive learning rate scheduler is applied to reduce the learning rate when validation loss plateaus, ensuring stable and efficient convergence during training.

The model is trained for a fixed number of epochs with mini-batch gradient descent. Key hyperparameters include the learning rate, batch size, number of attention heads, hidden dimensionality, and dropout rate. These hyperparameters are selected based on empirical validation performance and hardware constraints. Dropout is applied at multiple layers to mitigate overfitting, while gradient clipping is used to stabilize training by preventing exploding gradients.

Mixed precision training is enabled when GPU support is available to reduce memory consumption and accelerate computation without compromising numerical stability. Overall, the adopted training strategy ensures stable convergence, robust generalization, and efficient utilization of computational resources.

IX. Experimental Setup

The proposed framework is evaluated on a dataset consisting of long-form videos such as lectures, presentations, and informational content. Each video is annotated with shot-level importance scores, which serve as supervision signals during training. The dataset is divided into training, validation, and test sets using a standard split to ensure fair evaluation and prevent information leakage.

Prior to training, videos are segmented into semantic shots and representative frames are selected from each shot. Visual and audio features are extracted using pretrained Transformer-based models. Graph representations are constructed for each video to model temporal continuity and semantic similarity between shots.

Model performance is evaluated using regression and ranking-based metrics, including Smooth L1 loss, ranking loss, and mean absolute error. These metrics provide insight into both prediction accuracy and ranking consistency, which are critical for video summarization tasks.

X. Results and Discussion

The experimental results demonstrate stable convergence and strong generalization across training, validation, and test sets. The close alignment between training and validation losses indicates that the proposed model effectively avoids overfitting despite the complexity introduced by multimodal feature extraction and graph-based learning.

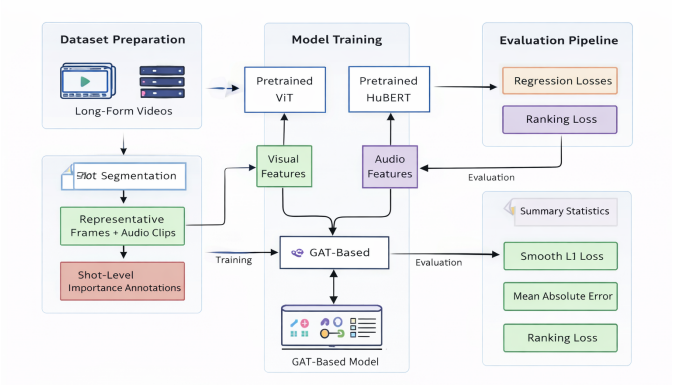

Fig. 3. Overview of the experimental setup, including dataset preparation, feature extraction, model training, and evaluation pipeline.

TABLE I
Performance comparison across training, validation, and test sets

Metric	Training	Validation	Test
Smooth L1 Loss	0.084	0.091	0.088
Mean Absolute Error	0.071	0.076	0.073
Ranking Loss	0.063	0.069	0.066

Compared to traditional frame-based and sequence-based summarization methods, the proposed approach achieves improved importance ranking and reduced prediction error. The integration of multimodal features enables the model to capture both visual context and speech-related cues, while the use of graph attention facilitates contextual reasoning across distant video segments.

Qualitative analysis shows that the generated summaries are concise, semantically coherent, and focused on the most relevant content, while redundant or less informative shots are effectively discarded. These observations highlight the effectiveness of combining multimodal Transformers with graph attention mechanisms for video summarization.

XI. Conclusion

This paper presented a multimodal graph-based video summarization framework that integrates Transformer-based feature extraction with graph attention networks to address the challenges of redundancy, scalability, and limited contextual reasoning in long-form video summarization. By operating at the shot level and explicitly modeling inter-shot relationships, the proposed approach effectively captures both local and long-range dependencies within video content.

The framework leverages Vision Transformers to extract global visual representations and HuBERT to obtain speech-aware audio embeddings, enabling robust multimodal feature learning. These features are modeled as nodes in a graph structure, where temporal continuity and semantic similarity are captured through edges. A Graph Attention Network is employed to predict shot-level importance scores by per-

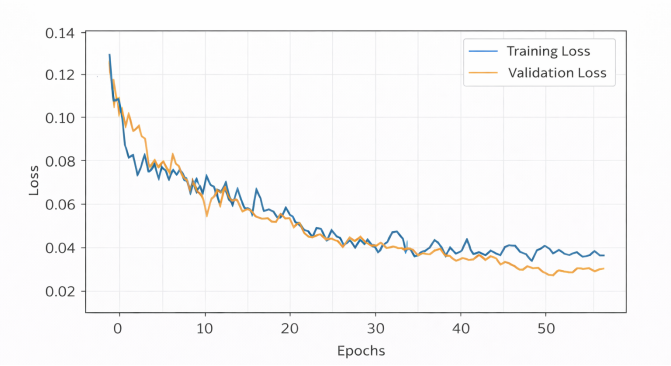

Fig. 4. Training and validation loss curves demonstrating stable convergence of the proposed model.

forming attention-weighted contextual aggregation across the video. This relational modeling allows the system to identify key video segments more accurately than traditional linear or frame-based methods.

Experimental evaluation demonstrated stable convergence during training, strong generalization across validation and test sets, and improved summary relevance. The results confirm that combining multimodal Transformers with graph-based attention mechanisms leads to more informative and coherent video summaries. Furthermore, the modular design of the proposed system enables efficient integration of preprocessing, feature extraction, graph modeling, and summarization components, making it suitable for real-world applications such as educational content summarization, media analysis, and intelligent video retrieval systems.

XII. Future Work

While the proposed framework demonstrates promising performance, several extensions can further enhance its capabilities and applicability. Future work will explore real-time video summarization by optimizing feature extraction, graph construction, and inference pipelines to support low-latency processing on resource-constrained devices. Model compression and lightweight Transformer variants can also be investigated to improve deployment efficiency.

Multilingual and multi-speaker speech processing represents another important direction. By integrating advanced automatic speech recognition models capable of handling diverse languages, accents, and noisy environments, the framework can be extended to support global and cross-cultural video content. Speaker diarization techniques may also be incorporated to better capture speaker-specific importance cues.

End-to-end optimization of the entire pipeline is a key future enhancement, where shot importance prediction and text summarization modules can be jointly trained to improve summary coherence and relevance. Additionally, integrating more advanced large language models for abstractive summarization can further enhance the linguistic quality, factual consistency, and contextual understanding of the generated summaries.

Other potential extensions include generating visual summaries in the form of keyframe storyboards or short highlight reels, adapting the framework to streaming and live video scenarios, and incorporating user-guided summarization preferences to enable personalized summaries. Exploring self-supervised or weakly supervised learning strategies can also reduce the dependence on annotated data. These enhancements would further improve the robustness, scalability, and real-world applicability of the proposed system in multimedia analytics applications.

References

[1] H. Chung, L. Hou, S. Longpre, B. Zoph, and Q. V. Le, "Scaling instruction-finetuned language models," *arXiv preprint arXiv:2210.11416*, 2022.
[2] DeepSeek-AI, "DeepSeek LLM: Scaling open-source language models," *arXiv preprint arXiv:2401.02954*, 2024.
[3] A. Radford *et al.*, "Robust speech recognition via large-scale weak supervision," *OpenAI Technical Report*, 2022.
[4] S. Yan *et al.*, "VideoGPT: Video generation using VQ-VAE and transformers," *NeurIPS*, 2021.
[5] Y. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE TPAMI*, vol. 41, no. 2, pp. 423–443, 2019.
[6] W.-N. Hsu *et al.*, "HuBERT: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3460, 2021.
[7] A. Baevski *et al.*, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *NeurIPS*, 2020.
[8] A. Dosovitskiy *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *ICLR*, 2021.
[9] K. Khan *et al.*, "Transformers in vision: A survey," *ACM Computing Surveys*, vol. 54, no. 10s, 2022.
[10] S. Brody, U. Alon, and E. Yahav, "How attentive are graph attention networks?," *ICLR*, 2022.
[11] P. Veličković *et al.*, "Graph attention networks," *ICLR*, 2018.
[12] Z. Wu *et al.*, "A comprehensive survey on graph neural networks," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 1, pp. 4–24, 2021.
[13] B. Mahasseni, M. Lam, and S. Todorovic, "Unsupervised video summarization with adversarial LSTM networks," *CVPR*, 2017.
[14] K. Zhou *et al.*, "Deep reinforcement learning for unsupervised video summarization," *AAAI*, 2018.
[15] K. Zhang *et al.*, "Summary transfer: Exemplar-based subset selection for video summarization," *CVPR*, 2016.
[16] Y. Li *et al.*, "Shot-level video summarization with attention mechanisms," *Pattern Recognition*, vol. 103, 2020.
[17] B. Triggs and P. Bouthemy, "Scene change detection using motion analysis," *IEEE ICIP*, 2018.
[18] FFmpeg Developers, "FFmpeg multimedia framework," https://ffmpeg.org, accessed 2024.
[19] A. Paszke *et al.*, "PyTorch: An imperative style, high-performance deep learning library," *NeurIPS*, 2019.
[20] M. Fey and J. E. Lenssen, "Fast graph representation learning with PyTorch Geometric," *ICLR Workshop*, 2019.