Received: 20 December 2025 | Revised: 23 July 2026 | Accepted: 10 September 2026 | Published online: 20 November 2026

RESEARCH ARTICLE

Journal of Computational and Cognitive Engineering
2026, Vol. 00(00) 1–17
DOI: 10.47852/bonviewJCCE62028849

BON VIEW PUBLISHING

Unsupervised Video Summarization via BiLSTM-Attention and Reconstruction-Based Scoring

Manjusha Jagannath Yeola[1,*] and Sunita S. Barve[1]

[1] *Computer Engineering Department of MIT Academy of Engineering, Savitribai Phule Pune University, India*

Abstract: The goal of video summarization is to automatically summarize a video to produce an informative summary by extracting the representative portion of the video. Recent unsupervised methods have decreased the reliance on manual annotations, but by effectively representing long-range temporal relationships while keeping computations tractable, this is a major challenge. In this paper, we propose a compact frame-based unsupervised video summarization framework based on bidirectional long short-term memory (BiLSTM) network and multi-head attention mechanism for improving the temporal feature representation and keyframe selection. Extensive experiments were performed on two benchmark datasets, SumMe and TVSum, using six different BiLSTM-based attention configurations. To explore the effect of the attention mechanisms and training samples, the performance was evaluated by F1-score, precision, and recall. Experimental results indicate that the configuration of BiLSTM–4 Heads consistently yielded the best trade-off with the highest F1-scores of 0.5896 on the SumMe dataset and 0.6721 on the TVSum dataset. The ablation study also shows that four-head attention configuration consistently outperforms the conventional attention configurations in terms of contextual representation. In general, the proposed framework proves that the integration of BiLSTM, multi-head attention, and reconstruction-driven scoring is a computational efficient and effective approach for fully unsupervised video summarization.

Keywords: multi-head attention, reconstruction learning, temporal feature representation, unsupervised learning, video summarization

1. Introduction

In today's digital era, with the proliferation of video content, intelligent systems that can handle and process vast amounts of visual information are essential. These systems are widely used in various industries, such as video surveillance, sports analytics, e-learning platforms, and content recommendation engines. The importance of technologies that efficiently and rapidly generate accurate, concise, and comprehensible summaries of long movies without any human interaction is increasing day by day [1].

Video summarization is a technique which extracts and removes parts from a video and produces a shorter video, but maintains the original content or conveyed information.

Most of the traditional methods for solving this problem have been based on supervised learning techniques that necessitate extensive annotated data collections for training the models. Still, the production of accurate reference data is a very laborious process, is affected by human bias, and cannot be easily expanded to different areas, and on the other hand, unsupervised video summarization has gained the popularity by incorporating heuristic strategies such as adversarial and reinforcement learning approaches [2].

Transformer-based architectures and multimodal learning frameworks have showed promising results in recent years, even though they are not as reliable as they claim to be, in terms of the ability to grasp long-range dependencies and contextual relevance in video summarization systems [3–6]. Recurrent neural networks (RNNs), long short-term memory (LSTM), and the most modern bidirectional LSTM (BiLSTM) are among these architectures and can encompass the temporal relationships in sequential data [7, 8]. BiLSTM, especially, has found extensive usage in sequential video tasks as it can combine both past and future context information [9]. Attention mechanisms, for example, the Transformer architecture, are used to complement these sequence models, that allow the model to prioritize different frames or features on the fly, thus enhancing both interpretability and accuracy [1]. Many recent studies and reviews in this field are summarized in a recent overview by Otani et al. [10] which covers important models, datasets, and new evaluation trends. AI summarization systems have become key to understanding and managing content effectively, as shown by recent studies [11].

A fully unsupervised video summarization framework is introduced that combines the BiLSTM's temporal power with the multi-head self-attention's contextual sensitivity. We define an importance scoring technique grounded on the reconstruction concept: a video frame is visualized with a set of high-dimensional features which are the output of the pretrained ResNet50 model. Furthermore, the frame feature sequence is processed by a

***Corresponding author:** Manjusha Jagannath Yeola, Computer Engineering Department of MIT Academy of Engineering, Savitribai Phule Pune University, India. Email: 2024080440003@mitaoe.ac.in

© The Author(s) 2026. Published by BON VIEW PUBLISHING PTE. LTD. This is an open access article under the CC BY License (https://creativecommons.org/licenses/by/4.0/).

BiLSTM-attention network. It attempts to duplicate the first frame's features, and the difference in reconstruction from the original features, using mean squared error (MSE), is an indicator of the importance of the frame. That is, the frames that the model struggles to reconstruct are those that contain more unique or key information and that have a higher likelihood of being chosen for the final summary.

We have a totally unsupervised system, which does not require any type of manual annotations. It is a highly effective technique for various kinds of video contents and also is computation efficient due to its compactness. We perform experimental evaluations of our approach on two standard benchmark datasets SumMe and TVSum that have multi-user-annotated videos. The extensive experiments confirm that our model also performs competitively with other models in terms of F1-score and shows that the model has the potential to be used in the real world.

The main contributions of this work are:

1) Fully unsupervised video summarization with compact integration of BiLSTM, multi-head attention, and frame importance estimation using reconstruction.
2) A detailed comparison of six different variants of the BiLSTM attention model on the benchmark SumMe and TVSum datasets in three different training sample sizes.
3) A study in which a moderate number of heads (4) for the attention configuration is shown to be a good compromise between contextual representation and summarization.
4) Implementing and testing the methods extensively on precision, recall, F1-score as well as statistical analysis and comparing with representative state-of-the-art methods.

On the benchmark SumME and TVSum datasets, this proposed method achieved an average score of 53.8% and 61.2% respectively thereby demonstrating performance that is comparable to various established unsupervised models like DR-DSN, Cycle-SUM, and VAE-GAN. In addition, it shows generalization to different video domains. Ablation studies and loss/accuracy convergence plots highlight the scalability and modularity of the framework, highlighting its suitability for real-world deployment at the edge, mobile, and streaming scenarios.

The rest of the paper follows: Section 2 summarizes literature of both supervised and unsupervised video summarization. Section 3 describes in detail our proposed approach. A brief description of the experimental setup and implementation is given in Section 4. The comparative analysis is presented in Section 5, which also includes the outcomes. Section 6 wraps up the paper and talks about the next steps in research.

2. Related Work

Over the last 20 years, video summarization has very much moved from heuristic-based summarization techniques that do not require supervision to deep learning-based techniques. In this section, supervised and unsupervised approaches are overviewed, with recent innovations showing the new direction of research in it.

2.1. Supervised video summarization

Supervised methods heavily depend on labeled datasets in which summaries created by humans are used as the final reference. In previous published papers, classification-based models are used to decide the most important frames based on the text. Such models hinged on manually engineered features and frequently could not generalize to different kinds of videos. Deep learning CNNs and RNNs have been able to model temporal relationships on video sequences. Temporal relationships on video sequences are improved by deep learning CNNs and RNNs. In Reference [12], a thorough taxonomy of video summarization techniques was given in which new trends, challenges, and breakthroughs were pointed out in the area of video summarization in different learning paradigms. A large number of summarization models are based on BiLSTM to model temporal dependency, while others have been proposed that rely on an attention only model that allows parallel processing and more effectively captures long-term dependency. The ability to capture long-range dependencies is well captured by attention mechanisms, especially transformer-based architectures [4].

For instance, the method suggested in Reference [13] includes a dual self-attention mechanism and a two-stream capsule network which models inter-frame relationships without using recurrent network structures and provides better frame selection. In this work, the supervised design of this approach reveals the potential of hybrid attention architectures for the task of learning summary-worthy content.

2.2. Unsupervised and weakly supervised approaches

The purpose of unsupervised summarization is to generate video summarization without any explicit labels, and it is based on the intrinsic structure of the video data. The unsupervised early approaches were mostly based on basic visual features, such as motion, saliency, and color histograms to determine the most salient frames [8]. Even though these ways were instinctive, they did not have the ability to take into account semantic or temporal information.

With the rise of deep learning, unsupervised models started to use sequence modeling techniques. Later, Yuan and Zhang [14] proposed a deep reinforcement learning approach that leverages shot-level semantic information, thereby finding that leveraging contextual information at a higher level can yield much better results for unsupervised summarization. In unsupervised settings, Yoon et al. [15] proposed an interpolation-enhanced DRL framework that not only covers more time but also captures more semantics in the summary than traditional reward-only frameworks.

Recently, attention mechanisms enhance contextual representation and frame-importance estimation in video [9]. Our proposed approach involves selecting the hard-to-regenerate keyframes using sparse reconstruction and attention, a concept very similar to the reconstruction-based importance we have in our model [16]. In recent years, several unsupervised methods have adopted the paradigm of self-supervised and contrastive learning to enhance the quality of the representation and avoid annotation [17–19].

To summarize surveillance videos in educational institutions, one such method proposed by Muhammad et al. [20] is to model activity attention scores using a deep CNN model, and then extract the keyframes from a Campus Activities Dataset (CAD). Summaries are further improved in terms of representativeness using inter-frame similarity for redundancy reduction. Several multimodal summarization frameworks that have used both audio-visual cues have proven to be effective for informative segment detection [21].

A recent study [22] developed for instance a multimodal feature extractor based on pyAudioAnalysis and OpenCV and tested several supervised classifiers with impressive results for Random

Forest and F1 score gains after the classification. The use of this specific framework as a real-time web service illustrates the possibility of using such systems for education, surveillance, and content archiving.

The recent progress of sequence modeling also shows the effectiveness of hybrid memory–attention designs. In some cases, like study by Yuan et al. [23] which applied xLSTM–XGBoost ensemble to the problem of forecasting highly volatile time series, LSTM variants improved with the addition of lightweight learners can help stabilize and generalize the model in non-stationary settings. While introduced for price prediction, this architecture also emphasizes the importance of modernized LSTM architecture for complex temporal reasoning, which inspires our work of BiLSTM–Attention design. Also, audio-visual fusion techniques have been shown to enhance the semantic understanding through fusion of complementary modalities like speech and visual features [24].

Likewise, for the video understanding task, we can see that Deep Learning-Driven Body-Sensing Game Action Recognition using MediaPipe and YOLO demonstrates the significance of attention-guided detection frameworks that can detect salient human actions in real time, emphasizing the need for strong feature localization and interpretability.

2.3. Reconstruction-based and representation learning methods

One of the major developments in unsupervised summarization is the use of reconstruction loss as a measure of significance. An important development in unsupervised video summarization is the use of reconstruction objectives to identify informative video content. Reconstruction-based methods use a selector–reconstructor framework to assess how well selected content represents the original video. Li et al. [16] proposed an iterative framework that alternately optimizes the selector and reconstructor using reconstruction objectives, closely related to the reconstruction-based importance estimation used in our model, and Li et al. [25] presented a progressive video summarization method, in which the important video content is progressively identified to produce improved video summaries [25]. Recent works like the one by Li et al. [16] proposed a simplified GAN (SGAN) summarization model that does not require complicated dual-discriminator networks, as it employs iterative selector optimization and reconstruction consistency. Liu et al. [26] proposed a reinforcement learning model with a 3D spatio-temporal *U*-Net encoder to capture both spatial and temporal relationships simultaneously for generating high-quality summaries in an unsupervised manner. The authors have suggested a reinforcement learning model along with a 3D spatio-temporal *U*-net to capture spatial and temporal video dynamics in an unsupervised summarization scenario. At the same time, transfer learning is incorporated to enhance the overall performance. Unsupervised and hybrid learning are also gaining significant attention in video summarization tasks in the real world, as seen from recent surveys [27].

Deep neural networks (DNNs) that have been pretrained on ImageNet, such as ResNet50 and VGG16, are the ones that are mostly used to get strong visual features from the raw video frames [28]. Basu et al. [9] proposed attention-based model WAnet that introduced frame-weighting mechanisms, which augment the temporal attention to identify the importance of the respective frames based on the notion of visual saliency and positional relevance. Multi-source feature fusion techniques have also been investigated recently to further improve frame representation in an unsupervised context, as it has been shown that this approach leads to better contextual understanding and summarization results [29]. Xu et al. [30] suggested MHSCNet, which is based on the hierarchical convolutional network and multimodal input, as well as shot-aware segmentation, greatly enhancing the contextual coherence and content relevance of the summaries. In addition, researchers have expanded XAI-SUM framework by adding entropy- and diversity-based attention explanation signals to various architectures such as CA-SUM, SUM-GDA, and VASNet. Also, explainable AI methods like heat maps help provide additional insights into the model's focus and decision-making processes. Recent frameworks have started to model the subjectivity that is present in video annotations produced by humans. One prominent method is multi-annotation attention model (MAAM) [31] that combines label aggregation and model training in a manner that considers the relevance scores per frame as hidden variables learned using the expectation-maximization algorithm. In addition, MAAM uses Vision Transformers (ViT) to represent the global temporal dependencies, making it an interesting solution for temporal modeling for summarization tasks compared to RNNs. Most of the existing methods lack the ability to capture local motion patterns between the neighboring frames, whereas temporal attention mechanisms are effective at capturing sequence-level dependencies. To tackle this, the MMAN framework [32] introduces a motion attention module (PMAM) without any parameters and introduces multi-head attention with object and scene awareness to improve video understanding by fusing local motion and global contextual information. It is designed without any parameters, allowing for efficient modeling that is not complicated by additional training parameters. Additionally, knowledge-aware and multimodal representation learning models have been proven to further improve contextual understanding and summarization performance [33].

2.4. Benchmark datasets

Several benchmark datasets have been largely accepted as the standard measure to assess summarization systems. SumMe has short videos generated by users and multiple ground truth summaries per video. TVSum is a set of web videos, organized by topics. The importance scores are rated by multiple users. These datasets are still the standard for benchmarking and comparison of supervised and unsupervised models.

2.5. Summary and research gaps

Although recent breakthroughs have elevated the quality of video summaries to a great extent, the currently available unsupervised models are still hindered by difficulties in temporal modeling, interpretability, and adaptability to varying video domains [34, 35]. In contrast to this, there has been relatively little research that has investigated fully unsupervised environments that access both sequence-level dependency (using BiLSTM) and contextual relevance (using attention mechanisms). Further, reconstruction-based approaches in conjunction with hybrid architectures have been scarcely investigated.

To overcome these issues and limitations, a hybrid model of BiLSTM-attention is proposed to reconstruct frame-level features, and then, reconstruction error is served as the frame importance signal. Our model is trained without supervision and tested on standard datasets.

3. Methodology

3.1. Frame extraction and feature encoding

First, each of the raw input videos is broken up into a set of static frames, uniformly sampled over the length of the video. These are the simplest form of visual presentation units for the summarization.

In order to capture the semantic content of each frame, the pretrained ResNet50 architecture, pretrained on the ImageNet dataset, is used. So, this model extracts 2048 dimensional deep feature vectors in its last by last layer, thus offering a rich spatial encoding of visual features.

In a more formal way, let f_i be the image frame and x_i its feature vector. This is the result of applying the vector sequence to the video, which gives the high-level semantic representation of the video, designated as the vector sequence X.

3.2. Temporal modeling using BiLSTM and multi-head attention

The frame features extracted from the sentences are temporally structured, and this structure needs to be captured through temporal dependencies, which can only be modeled with a sequence. To encode both past and future contextual cues, we use a bidirectional long short-term memory (BiLSTM) network.

Let the concatenated forward and backward hidden states at time t as h_t. These representations are in turn refined in a multi-head attention component, which calculates multiple self-attention scores simultaneously. In turn, each attention head learns the inter-frame relationships by re-weighting features based on the temporal relevance, producing contextual embeddings better suited for downstream reconstruction.

Multi-head attention is composed of 4 heads, each of size 64, total size 256. The output of BiLSTM is projected linearly to a query, key, and value. Concatenation followed by a linear layer is used for head fusion. Because BiLSTM maintains the order of the frames, explicit position encoding was not used; ablation results indicated that sinusoidal encoding did not show any improvement.

3.3. Frame reconstruction via time distributed dense layer

In order to measure the importance of the frames in an unsupervised manner, we generate the features of the original frames from the attention-augmented representations through a time-distributed fully connected layer. Let $\hat{x}_t$ be the reconstructed feature of the frame t. The training objective of the model is to reduce the reconstruction error that is given by the mean squared error (MSE) as defined in Equation (1):

$$L_{recon} = \frac{1}{T}\sum_{t=1}^{T} ||x_t - \hat{x}_t||_2^{\,2} \tag{1}$$

where x_t is the original feature and $\hat{x}_t$ its reconstruction.

The loss function here acts as a self-supervised signal, which allows the model to learn feature transformations that recognize important temporal patterns and retain semantic fidelity.

3.4. Frame importance scoring

Choosing the right frame is a key factor in deciding the quality and the extent to which the summary can be understood visually. The significance of a frame is determined in an objective way via a reconstruction-based scoring criterion, which is based on how difficult it is for the model to reconstruct each frame from its latent representation, in the proposed framework. Let's take a look at an example of a content feature vector and its reconstruction. The reconstruction error along with the content feature vector can be used as a representative of semantic novelty and temporal distinctiveness.

The interpretation of those frames, which results in larger reconstruction error, is that they contain unique or unpredictable content that cannot be derived from the surrounding frames—considering this, therefore, they are considered as informative ones.

The opposite, however, is true of frames with low reconstruction errors; if such frames occur that means that the information is redundant or that the content is contextually repetitive and that the parts associated with those images are not represented in the summary. Implicit importance values are min-max normalized to the range [0, 1] so as to be consistent with videos of different lengths and complexities.

This is an interpretable reconstruction-based scoring strategy, on which the subsequent frame selection process is based.

The model keeps only part of the frames that represent the maximum semantic variance and maximum narrative value, by ordering frames according to their normalized reconstruction errors. Unlike heuristic saliency or motion based scoring, it is data-driven, self-supervised, and temporally aware, allowing the system to select the keyframes without relying on external supervision and still maintain the temporal continuity between the adjacent frames.

3.5. Summary generation via KTS and Knapsack optimization

Finally, we use a segment-level selection approach that combines kernel temporal segmentation (KTS) and Knapsack optimization to generate video summary. This technique creates an accurate summary, but retains the most semantically relevant and meaningful content.

KTS is applied to the frame-level features extracted from the video using ResNet50, and the self-similarity matrix is localized to temporally coherent segments, which is the first step. As a result, each segment can be viewed as a visually consistent picture or action. We compute for each segment:

1) Average importance score (mean of normalized reconstruction-based scores).
2) The length (frames or seconds).

We then create a 0–1 Knapsack problem, where each segment is considered an item with:

3) Value = average importance score
4) Weight = segment duration.

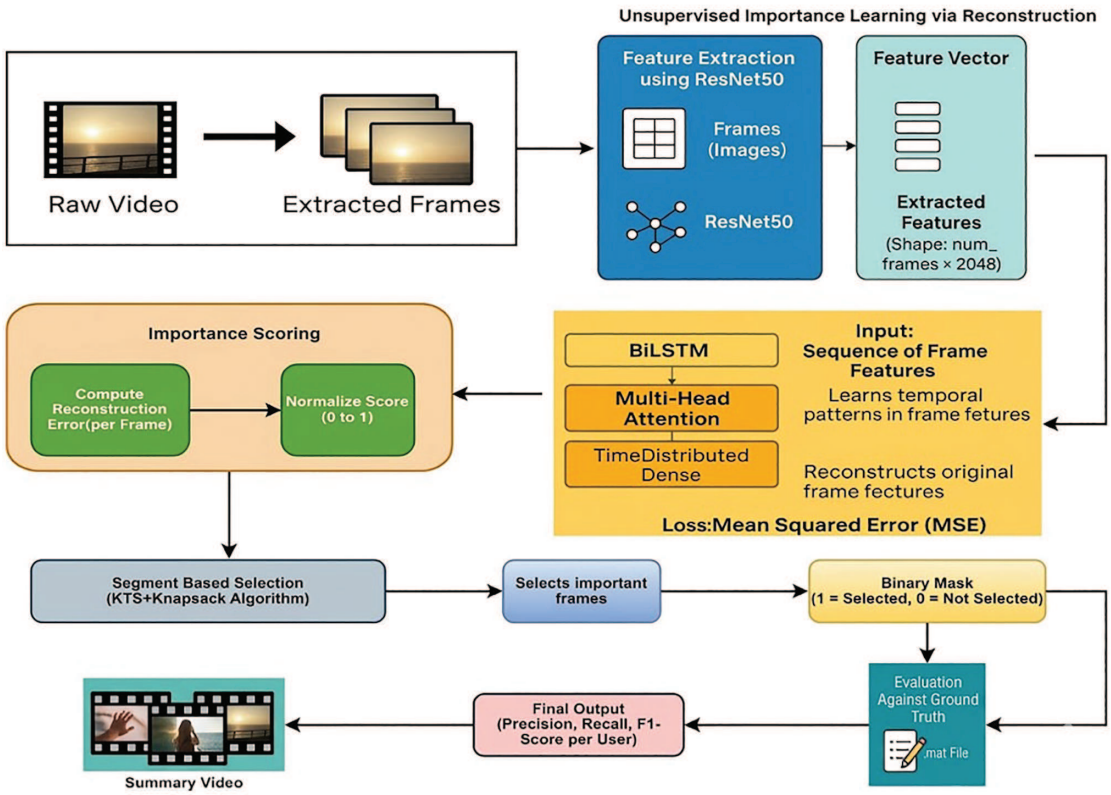
The objective is to select a set of non-overlapping segments with a total duration of up to 15% of the original video length while the total importance score is as high as possible.

The selected segments are then concatenated temporally, to generate the final video summary. This KTS-Knapsack strategy is a principled, interpretable approach respecting scene segments and focusing on semantically rich content.

3.6. System architecture overview

The proposed video summarizing system with no supervision is given in Figure 1. The system is a set of tightly coupled modules

Figure 1
System architecture



that work together to implement frame-level importance estimation completely unsupervised. The pipeline initially samples at a fixed rate in the raw video. The following study uses these frames as the basic visual primitives.

Then, each frame is input to a ResNet50 convolutional neural network which was pretrained on the ImageNet data set. This model was selected because it is very effective in generating feature embeddings that are semantically informative, discriminant, and efficient. Each frame is represented as 2048 dimensional vector, and the output of this step is a set of deep visual features. These are the feature vectors that are fed into the central learning module of the system.

The reconstructed features are provided by the time distributed dense layer. Here, the emphasis will be on the ResNet50 features that are originally extracted for each frame, and the goal will be to apply the reconstruction constraint at this level. At this stage, the system is performing an approximation of the input sequence of features using its latent representation. The quality of reconstruction is measured with mean squared error (MSE) and per-frame reconstruction error based on it is taken as the measure of the importance of a frame.

The reasoning is that those frames which result in more significant reconstruction errors are the ones that contains more distinctive visual information; thus, their informativeness is higher, as much as they are the most difficult ones to recreate.

In the last summarization step, the feature sequence of the segmented video is processed by kernel temporal segmentation (KTS) and Knapsack algorithm. Average importance score per segment is computed, and its second count per segment is taken from KTS. The Knapsack optimization is then used to find the segments that sum up in length to not exceed 15% of the length of the video. Maximizing the sum of the normalized importance scores pushes the selection of the segments towards picking up the most interesting parts of the video to be represented by the selected fragments.

The frames are selected based on percentile or fixed ratio selection (e.g., top 15% of top-scoring frames). A binary mask is generated to denote which frames are kept or discarded, and the frames selected are then joined in their original time sequence to make the summary video. The quality of the generated summary is quantified by comparing the binary mask with the user-annotated ground truth information in benchmark datasets using precision, recall, and F1 score.

The proposed framework is modular and can be fully unsupervised, as it takes advantage of deep visual feature encoding, sequential modeling, attention-based context refinement, and reconstruction-driven importance estimation.

3.7. Mathematical formulation of the summarization objective

Let a video V be represented as a sequence of frame features $X = \{x1, x2, \ldots, x_T\}$ where each $x_t \in R^d$ is extracted using a pretrained ResNet50.

We encode the temporal context as defined in the Equation (2) in the BiLSTM-Attention model:

$$h_t = BiLSTM(x_t)\ \ a_t = MultiHeadAttention(h_t) \qquad (2)$$

The shared time distributed dense network $\hat{x}_t = \emptyset(a_t)$ is used for the reconstruction of the model.

An importance score is computed based on the reconstruction error in the Equation (3):

$$s_t = ||x_t - \hat{x}_t||_2 \Rightarrow \bar{s}_t = \frac{s_t - \min(s)}{\max(s) - \min(s)} \tag{3}$$

Using a binary mask as in eq. 4, frames are selected for the summary:

$$m_t = \begin{cases} 1, & if\ \bar{s}_t \geq 0 \\ 0, & otherwise \end{cases} \tag{4}$$

The overall model learns to minimize the reconstruction loss as shown below:

$$L_{recon} = \frac{1}{T}\sum_{t=1}^{T} ||x_t - \hat{x}_t||_2^{\ 2}$$

To get the summary, KTS divides the video into N segments $\{seg_1, seg_2, \ldots.., seg_N\}$, each with:

v_i: Average normalized importance score of segments i

w_i: Duration of segment iii in frames

W: Maximum allowed duration of the summary (typically 15% of total video length)

We introduce a binary decision variable $x_i \epsilon \{0, 1\}$ for each i denoting selection. The Knapsack objective is given in eq. 5 with the constraint defined in Equation (6):

$$maximize \sum_{i=1}^{N} v_i * x_i \tag{5}$$

$$subject\ to \sum_{i=1}^{N} w_i * x_i \leq W \tag{6}$$

where $x_i \epsilon \{0, 1\}, \forall_i = 1 \ldots..N$

This formulation guarantees that all the video segments that are selected are the most informative and diverse and that the summary length constraint is fulfilled.

Algorithm 1: Unsupervised Frame Importance Learning via Reconstruction

Input: Video V with frames $\{f_1, f_2, \ldots\ldots, f_T\}$
Output: Summary S consisting of selected keyframes
1. Extract features: $x_t < -ResNet50(f_t)\ for\ all\ t = 1\ to\ T$
2. Encode temporal context: $h_t < -BiLSTM(x_t)$
3. Apply multi-head attention: $a_t < -Attention(h_t)$
4. Reconstruct features: $\hat{x}_t < -Dense(a_t)$
5. Compute importance score: $s_t < -||x_t - \hat{x}_t||^2$
6. Normalize scores: $s_t < -MinMax(s_t)$
7. Apply KTS on $\{x_t\}$ to obtain temporal segments $\{seg_1, seg_2, \ldots.., seg_N\}$
8. For each seg_i, compute average score $\bar{s}_i$ and duration d_i
9. Apply Knapsack algorithm to select subset of segments covering $\approx$ 15% of video duration with maximum $\sum \bar{s}_i$

4. Experimental Setup

The experimental setup in this part was aimed at assessing the proposed unsupervised video summarization framework. Here, we present the datasets, training parameters, implementation tools, and evaluation protocol implemented in this research.

4.1. Datasets

We test our model on two standard video summarization datasets: SumMe [36, 37] and TVSum [38].

SumMe: A set of 25 videos which were created by users to demonstrate various real-life scenarios. Several users have annotated each video to create the ground truth for video evaluation. The average length of the videos is around 2 minutes.

There are 50 videos in TVSum, which were divided into 10 semantic categories (e.g., news videos, sports videos, cooking videos) from YouTube. 20 users rate each video with an importance rating of 1-5.

Both datasets are a common ground-truth for video summarization research and are equipped with user-annotated data for making objective performance comparison.

4.2. Preprocessing and feature extraction

To maintain consistency, rather than a uniform frame rate, we propose to sample a fixed number of frames per video when using videos of different durations. Therefore, each video has a number of frames sampled: 750, 1750, or 3000 frames per video, or approximately 3, 8, or 14 frames per second (fps). The changes are done across three different training regimens, for a total of 15k, 35k, and 60k frames in the dataset.

The size of the sampled frame is varied, and the frame is passed through a trained ImageNet-based ResNet50 model. Then, a 2048-dimensional feature vector is generated by the global average pooling layer. This series of deep feature embeddings is the input to the BiLSTM–Attention model for temporal modeling and importance estimation.

Only BiLSTM–Attention and reconstruction layers were trained, while the ResNet50 backbone was treated as a frozen feature extractor trained with ImageNet. Adam optimizer was used for training with learning rate 1×10^{-5}, which resulted from a grid search in the range [10^{-4}, 10^{-6}] for stable convergence with longer sequences. A small batch size of two was adopted because of the memory limitation of the GPUs for processing full-length video sequences—similar to previous summarization work. All experiments were performed using fixed random seeds {42, 123, 2023, 7, 99}, and the mean ± standard deviation over five runs is reported. Gradient clipping was used at 1.0, and early stopping (patience of 8) was used to avoid overfitting.

4.3. Evaluation protocol

The proposed framework was tested according to the precision, recall, and F1-score, which are the temporal overlap between the generated summaries and the user-annotated reference summaries. After applying KTS and Knapsack optimization, the predicted summaries were obtained for the selected frames, and ground-truth summaries were extracted from user-assigned importance annotations on the frames. Each video was evaluated with an F1 score calculated based on the temporal intersection-over-union (IoU), and the results reported are the average scores over all videos.

The model was trained in a self-supervised fashion, using reconstruction loss on the full dataset (including the assessed video), with no user annotations, frame-level importance scores, or summary labels used during optimization. No per-video fine-tuning or adaptation at test-time was done, and the same model was used for all videos in both datasets. Only at the end of evaluation, ground-truth annotations are used to calculate the precision,

recall, and F1-score. As a consequence, the adopted protocol is based on a transductive self-supervised evaluation scheme and not on a typical train/test partitioning scheme.

4.4. Summary of setup

The key experimental settings and model parameters used in this study are summarized in Table 1

Table 1
Model parameters

Parameter	Value
Datasets	SumMe, TVSum
Frame Sampling	~3 fps (750 frames/video), ~8 fps (1750), ~14 fps (3000)
Feature Extractor	ResNet50
Feature Dim	2048
Optimizer	Adam
Learning Rate	0.00001
Batch Size	2
Epochs	50
Loss Function	MSE
Evaluation Metric	F1-score
Tools	TensorFlow, Keras, Python
Hardware	AMD Ryzen 7 Processor, 24GB RAM, NVIDIA RTX 4050 GPU

5. Results and Discussion

5.1. Comparative evaluation protocol

In this subsection, the experimental procedure followed for the evaluation of the proposed BiLSTM–Attention framework with various architectural structures and training methods is described. All experiments were conducted with the same feature extraction and optimization parameters and evaluation metric, with the only difference being the use of multi-head attention in the evaluation. This was done by evaluating the performance with the benchmark SumMe and TVSum datasets in terms of F1-score, precision, and recall to provide a comprehensive comparison of the quality of the produced summaries and the retrieval effectiveness.

The attention configurations are used in conjunction with model variants. To understand the contribution of different attention mechanisms, six BiLSTM-based architectures namely BiLSTM No Attention, BiLSTM Soft Attention, BiLSTM with 2-Head Attention, BiLSTM with 4-Head Attention (Proposed), BiLSTM with 8-Head Attention, and BiLSTM with variational autoencoder (VAE) are tested under the same experimental setups. To ensure that the performance differences are not due to differences in the visual feature representation, all variants shared the same pretrained ResNet50 backbones for extracting frame-level features of 2048 dimensions.

1) Train and scale data

For testing the robustness and scalability, experiments were carried out with the three training regimes of about 15K, 35K, and 60K sampled frames. The settings are low, medium, and high volume training, respectively, allowing to assess model behavior in varying data availability settings. The same preprocessing pipeline, optimizer configuration, learning rate, batch size, and random seed initialization were maintained throughout all experiments to ensure a fair comparison across model variants and sample sizes.

2) Evaluation metrics

The standard video summarization metrics of precision, recall, and F1-score were used to assess performance, which measures the level of agreement between the automatically generated summary and the ground truth annotated by the user. Precision is the percentage of keyframes that are correctly selected from all the keyframes predicted, Recall is the percentage of keyframes that are correctly retrieved from all the keyframes that are relevant, and the F1 score is a balanced average of the two. For each experimental setup, these metrics have been calculated on both the benchmark datasets, according to the evaluation protocol from Section 4.

3) Experimental objectives

The experimental evaluation aimed at: (i) evaluating the impact of the various attention mechanisms that were introduced on summarization performance, (ii) studying the effect of the number of training samples on robustness and generalization of the models, (iii) identifying which attention-head configuration is the best for the reconstruction-based unsupervised summarization, and (iv) comparing the proposed BiLSTM–4 Heads architecture with other attention heads of the BiLSTM model in a consistent quantitative evaluation. This protocol is used to give a systematic basis for the detailed performance analysis in the following parts.

5.2. Quantitative results and performance analysis

A comprehensive quantitative assessment was conducted to assess the efficiency of the BiLSTM–Attention model, as the experimental setup had been followed, on the SumMe and TVSum data sets.

The evaluation metric included the mean of F1-score, Precision, and Recall that were averaged over five independent runs for each attention-head configuration (0, 2, 4, and 8 heads) and three sizes of training sample (15 k, 30 k, and 60 k). The combined results are presented and analyzed in the following subsections.

5.2.1. Performance overview

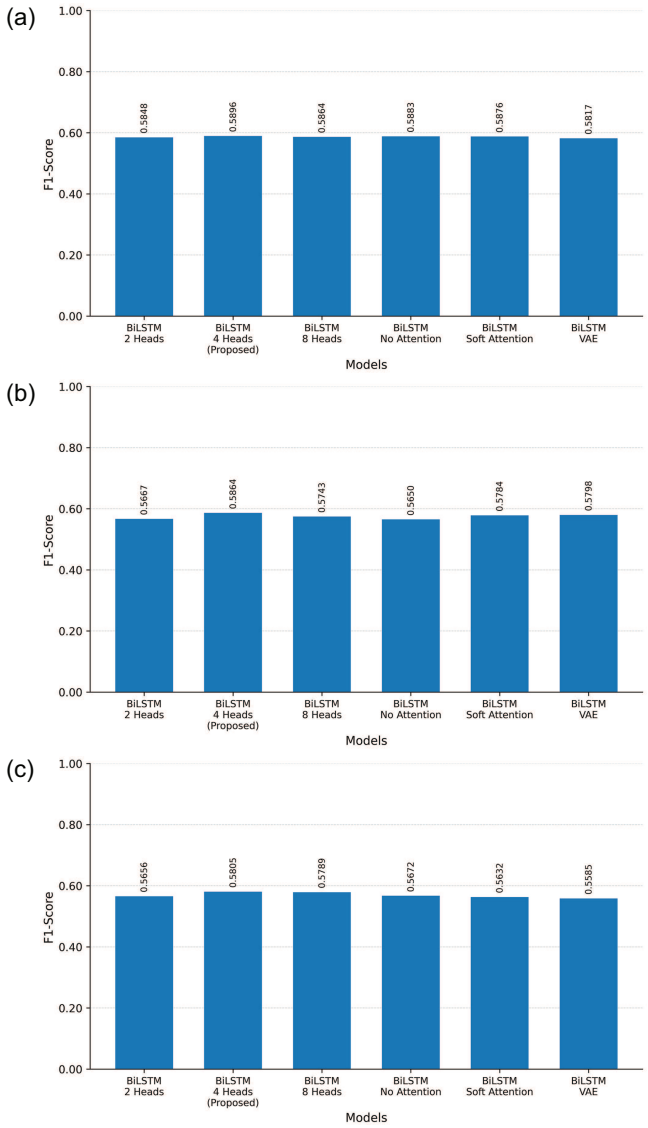
The proposed BiLSTM–4 Heads architecture was tested with five variants of BiLSTM on two benchmark datasets SumMe and TVSum using three different training regimes of 15K, 35K, and 60K sampled frames. The results on both the datasets showed that the proposed configuration consistently obtained one of the highest F1-scores, with a good balance between the precision and recall. On SumMe, the proposed model obtained F1-scores of 0.5896, 0.5864, and 0.5805 for the 15K, 35K, and 60K settings, respectively. Likewise, on TVSum, it had F1-scores of 0.6721, 0.6659, and 0.6684 under the respective trainings. The results show that using a moderate four-head attention configuration results in slight but stable gains compared to other variants of BiLSTM for summarization task, and also slightly better results than the other configurations across the different training sample sizes.

5.2.2. Comparison across attention-head configurations

Performance of the models evaluated on the SumMe dataset on varying training sample sizes is shown in the comparative bar plots of Figure 2 (a)–(c). The proposed BiLSTM–4 Heads configuration always obtained the best-F1 score, showing a good balance between the precision and recall. The improvements from the rest of the variants are not significant, but the four-head variant showed more stable summarization performance than the conventional BiLSTM, Soft Attention, VAE-based, and higher-head architectures [10, 20].

Overall, the obtained results indicate that a moderate four-head attention configuration is a promising trade-off between the representation of contextual features and computational complexity. This shows that more attention heads were not always better for summarization, suggesting that adding more heads might result in feature interactions that are redundant, but do not provide performance benefits that scale proportionally.

Figure 2
F1-score comparison of different BiLSTM variants on the SumMe dataset using approximately (a) 15K, (b) 35K, and (c) 60K sampled frames



5.2.3. Precision–recall behavior

Figures 3 and 4 show the comparative precision–recall performance of the evaluated video summarization techniques based on BiLSTMs with SumMe and TVSum data sets, respectively. Plots are used to offer complementary evaluation of summarization quality, showing the trade-off between recall and precision for various model setups.

The proposed BiLSTM–4 Heads configuration resulted in a well-balanced precision and recall with SumMe dataset, showing the balanced summarization over a variety of videos with different temporal properties. The proposed model also outperformed the other BiLSTM models in terms of recall in comparison with the other BiLSTM models on the TVSum dataset, while still being competitive in terms of Precision.

The precision–recall analysis shows that there is a moderate trade-off between compactness of the summarization and preservation of content in the four-head attention configuration. The performance gap between the variants evaluated is quite small; however, the proposed architecture always exhibits stable behavior on both test sets without being influenced by the test data, introducing unnecessary model complexity.

5.2.4. Influence of training-sample size

The F1-scores of the test models (different variants of BiLSTM) are shown as the function of different number of training frames (approx. 15K, 35K, 60K sampled frames) in Figures 5(a) and (b) for TVSum and SumMe, respectively.

1) TVSum Dataset

The model with the highest F1 score was the proposed BiLSTM–4 Heads model with about 15K sampled frames with an F1 score of 67.21%, followed by 66.59% for 35K training frames and 66.84% for 60K training frames. The same behavior was seen for other BiLSTM architectures where the performance change was only slight with the number of training samples but not a gradual improvement. The results show that the proposed architecture is able to learn the temporal dependencies even with relatively smaller training sets.

2) SumMe Dataset

The same pattern was seen with the SumMe data set. The proposed BiLSTM–4 Heads model obtained F1-scores of 58.96%, 58.64%, and 58.05% for the 15K, 35K, and 60K settings, respectively. There were some minor changes in performance for various attention configurations, but the ranking of the proposed model did not change much. This behavior indicates that the more frames are sampled the greater the amount of summarization does not necessarily correspond to better performance on video collections with greater diversity.

The key observations are as follows:

The proposed BiLSTM–4 Heads architecture has always got one of the highest F1-scores among the three training schemes on both benchmark datasets.

When the number of training examples was increased beyond 15K, there were relatively small changes in the performance of the proposed framework, suggesting that the proposed framework is fairly robust against the changes in the size of training set.

Figure 3
Precision–recall comparison of the evaluated BiLSTM variants on the SumMe dataset

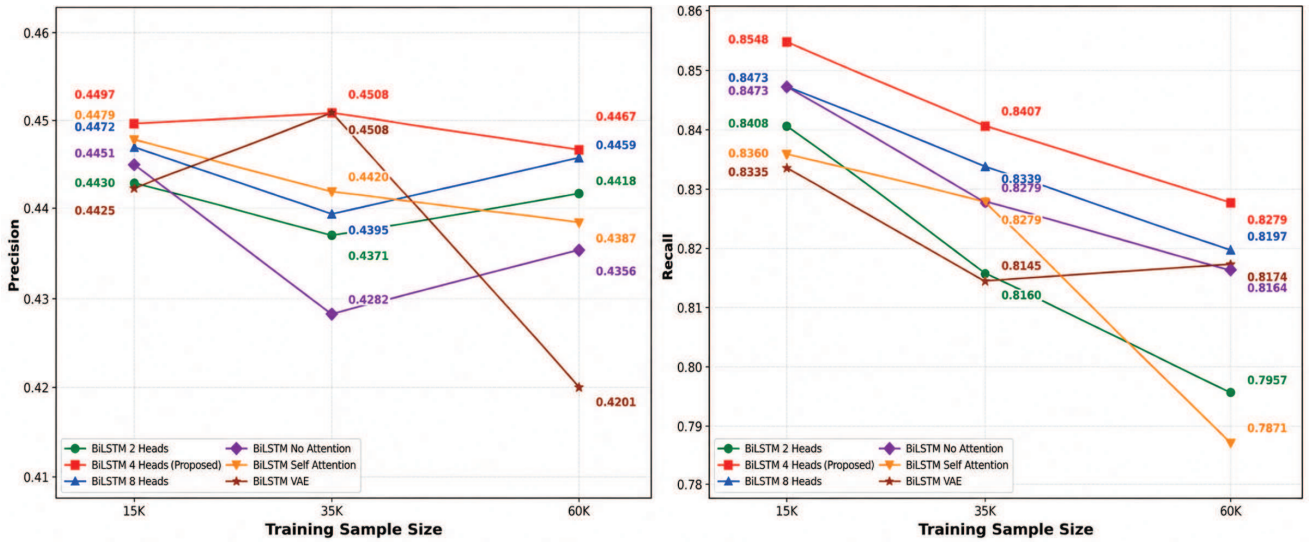
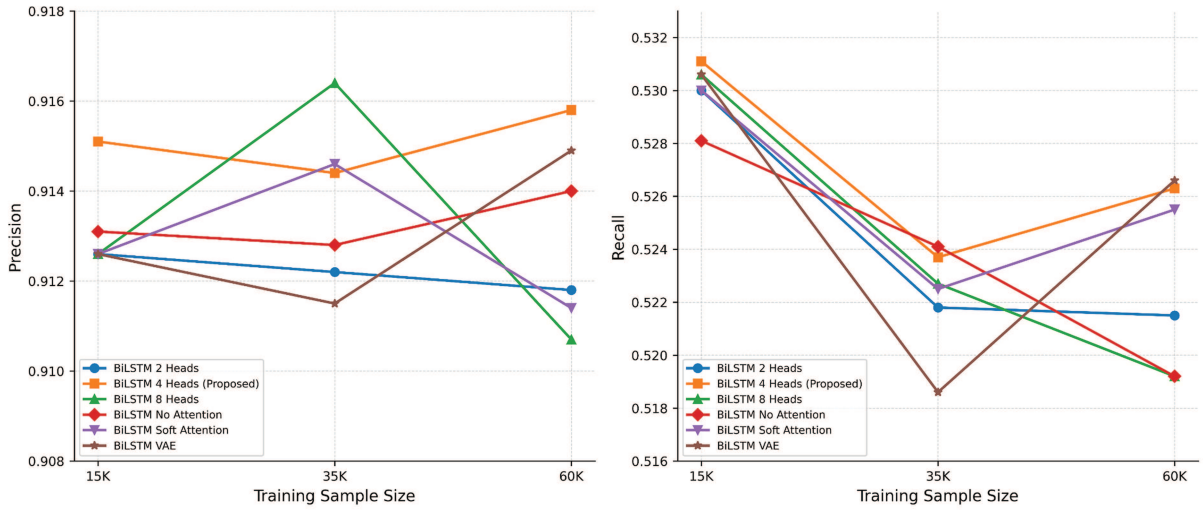


Figure 4
Precision–recall comparison of the evaluated BiLSTM variants on the TVSum dataset



The results indicate that the attention configuration has a more important effect on summarization accuracy than just adding more of the sampled training frames.

5.2.5. Dataset-specific observations

In absolute terms, TVSum fares better than SumMe in terms of F1-scores, which points to the fact that the video content is more structured in TVSum and less variable over time. TVSum performs better than SumMe in terms of absolute F1-scores, while showing similar trends of performance across the different explored BiLSTMs; this could be attributed to the more structured nature of the video content in TVSum and the lesser temporal variability. However, the proposed BiLSTM–4 Heads configuration has always been one of the top-ranked F1-scores on both datasets. The performance gap compared to the rest of the variants is generally very small and stable, suggesting that attention setup in an appropriate way helps to stabilize summarization performance for different video types.

5.2.6. Summary of quantitative insights

The proposed BiLSTM–4 Heads configuration has consistently obtained one of the best F1 scores on both the benchmark datasets.

Use of four-head attention is effective for the temporal representation, which is a good balance between the number of features and complexity of the model.

The performance changes were small when the training sample size increased beyond 15K, suggesting that the attention configuration is more important than the amount of training data for summarization performance.

In summary, the experimental results show the reproducibility and consistency of the experimental results under various attention settings and benchmark sets.

The quantitative results are used as a basis for the comparative analysis in the next section.

5.2.7. Statistical significance analysis

The per-video average F1 scores were computed for each video of the SumMe and TVSum datasets and then compared statistically. The proposed BiLSTM–4 Heads model was compared with the variants of the BiLSTM–No Attention and BiLSTM–2 Heads. A paired observation is thus the average F1-score of the same video with two model configurations. The Shapiro–Wilk test was used to determine normality of the paired differences. The normality assumption was not met, so the Wilcoxon signed-rank test was used as the main test, and the paired *t*-test was also reported for completeness. This procedure was performed for all four pairwise comparisons, using Holm correction.

As can be seen in Table 2, modest descriptive improvements were found for SumMe but none of the differences between the models were statistically significant at a $p < 0.05$ level. The per video average F1 distributions for the compared configurations were almost the same on TVSum and were not found to be statistically different. The effect sizes were all very small, so that the differences in performance should be considered relatively small rather than significant.

5.3. Comparative discussion of attention configurations

The experimental results show that the BiLSTM-based summarization framework with multi-head attention consistently outperforms the baseline models on both summarization datasets. The proposed BiLSTM–4 Heads model showed one of the best F1-scores and a good balance of precision and recall for various training methods (Figures 5 (a) and (b)), among the considered configurations. Such behavior indicates that a medium number of attention heads is a better approximation of temporal dependencies than is either a low number of attention heads or a high number of attention heads.

The attention mechanism is used to give different weights to temporal features when modeling a sequence, unlike a normal BiLSTM. This means, the model is more capable of highlighting informative video segments and retaining the long-range contextual information. The experimental findings suggest the advantage of the four-head configuration over representational capacity and computational complexity, while the number of heads beyond four is not always beneficial.

5.3.1. Diminishing returns at higher attention head counts

The results of the four attention heads and more than four attention heads did not consistently show improved summarization performance on either dataset. As presented in Tables 3 and 4, the BiLSTM–8 Heads configuration achieved performance similar to, but lower than, that of the proposed model, BiLSTM–4 Heads. On the SumMe dataset, the four-head model outperformed the others with all three training regimes, while on TVSum the differences between the four-head, eight-head, and VAE-based versions were not significant.

One likely reason is that the more attention heads increase, the more representational complexity the system must account for, but the more contextual modeling that is gained. As a result, the feature space available is spread out across more attention subspaces, which could make each of the attention mechanisms less effective when learning from relatively small datasets of

Figure 5
F1-score comparison across different training sample sizes: (a) TVSum dataset and (b) SumMe dataset

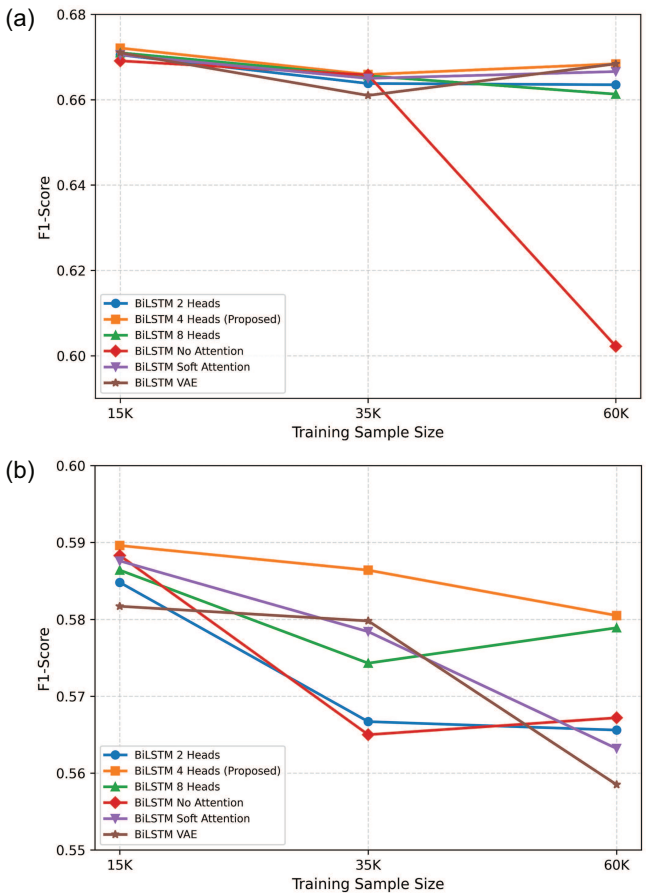


Table 2
Statistical tests

Dataset	Comparison	Shapiro–Wilk *p*	Paired *t*-test *p*	Wilcoxon *p*	Holm-adjusted *p*	Cohen's *d*
SumMe	No Attention vs. 4 H	<0.001	0.113	0.093	0.372	0.34
SumMe	2 H vs. 4 H	<0.001	0.863	0.715	0.952	0.04
TVSum	No Attention vs. 4 H	<0.001	0.224	0.465	0.952	–0.17
TVSum	2 H vs.4 H	<0.001	0.322	0.317	0.952	–0.14

Table 3
SumMe attention model performance

Sample Size	Model	F1-Score	Precision	Recall
15k	BiLSTM-2 Heads	0.5848	0.4466	0.8470
	BiLSTM-4 Heads (Proposed)	0.5896	0.4499	0.8553
	BiLSTM-8 Heads	0.5864	0.4473	0.8511
	BiLSTM No Attention	0.5883	0.4484	0.8553
	BiLSTM Soft Attention	0.5876	0.4499	0.8470
	BiLSTM VAE	0.5817	0.4430	0.8470
35k	BiLSTM-2 Heads	0.5667	0.4341	0.8161
	BiLSTM-4 Heads (Proposed)	0.5864	0.4491	0.8449
	BiLSTM-8 Heads	0.5743	0.4370	0.8375
	BiLSTM No Attention	0.5650	0.4282	0.8305
	BiLSTM Soft Attention	0.5784	0.4428	0.8340
	BiLSTM VAE	0.5798	0.4495	0.8168
60k	BiLSTM-2 Heads	0.5656	0.4383	0.7975
	BiLSTM-4 Heads (Proposed)	0.5805	0.4476	0.8257
	BiLSTM-8 Heads	0.5789	0.4440	0.8319
	BiLSTM No Attention	0.5672	0.4343	0.8175
	BiLSTM Soft Attention	0.5632	0.4373	0.7911
	BiLSTM VAE	0.5585	0.42402	0.8180

Table 4
TVSum attention model performance

Sample Size	Model	F1-Score	Precision	Recall
15k	BiLSTM-2 Heads	0.6705	0.9126	0.5300
	BiLSTM-4 Heads (Proposed)	0.6721	0.9151	0.5311
	BiLSTM-8 Heads	0.6710	0.9126	0.5306
	BiLSTM No Attention	0.6691	0.9131	0.5281
	BiLSTM Soft Attention	0.6705	0.9126	0.5300
	BiLSTM VAE	0.6710	0.9126	0.5306
35k	BiLSTM-2 Heads	0.6638	0.9122	0.5218
	BiLSTM-4 Heads (Proposed)	0.6659	0.9144	0.5237
	BiLSTM-8 Heads	0.6656	0.9164	0.5227
	BiLSTM No Attention	0.6658	0.9128	0.5241
	BiLSTM Soft Attention	0.6650	0.9146	0.5225
	BiLSTM VAE	0.6610	0.9115	0.5186
60k	BiLSTM-2 Heads	0.6635	0.9118	0.5215
	BiLSTM-4 Heads (Proposed)	0.6684	0.9158	0.5263
	BiLSTM-8 Heads	0.6613	0.9107	0.5192
	BiLSTM No Attention	0.6022	0.9140	0.5192
	BiLSTM Soft Attention	0.6666	0.9114	0.5255
	BiLSTM VAE	0.6684	0.9149	0.5266

benchmarks. Moderate attention configurations have been found to offer a good balance between representational power and computational efficiency in previous attention-based sequence modeling experiments [31, 39]. Multi-head attention captures diverse relationships across representation subspaces [40]. The results show that adding more heads at the attention does not necessarily reduce the performance of video summarization.

5.3.2. Generalizability across datasets

The key result from both SumMe and TVSum is that even though the datasets are structurally different, the relative improvement of attention, particularly with 4 heads, is consistent. SumMe consists of short, longer, and more varied videos, subjectively annotated, while TVSum consists of longer, thematically grouped videos with greater temporal coherence. The 4-head model resulted in slight but stable gains in both F1-scores of the two datasets, however, highlighting its transferability in the following aspects: it adapts to various forms of annotations, content domains, and video length (Figure 5 (a) and (b)). This is proof that the suggested architecture is capable of generalization for various summarization scenarios and is not overly tailored to the specific biases of the datasets.

5.4. Ablation study on attention heads and sample size

In order to see how the performance of video summarization is affected by the configuration of the attention head and the number of training samples, an ablation study was carried out. We considered two important factors: (*i*) the number of heads of attention to be used in the proposed BiLSTM–Attention architecture, and (ii) the impact of training schemes based on the number of sampled frames used to train the model (around 15K, 35K, and 60K). The acquired results give clues about the relationship between the attention configuration, the amount of training data, and summarization performance over the SumMe and TVSum benchmark datasets.

5.4.1. Effect of attention head count

The performance in summarization using different attention-head configurations (No Attention, 2, 4, and 8 Heads) is summarized as shown in Tables 3 and 4 and shown as shown in Figures 5 (a) and (b). However, in both SumMe and TVSum, multi-head attention, on average, brought only slight, but reliable, gains over the baseline BiLSTM model, suggesting that attention mechanisms contribute to the representation of temporal features for keyframe selection. The proposed BiLSTM–4 Heads model consistently performed one of the best performances on the SumMe and TVSum tasks under the three settings of 15K, 35K, and 60K learning, with 58.96%, 58.64%, 58.05% on SumMe and 67.21%, 66.59%, 66.84% on TVSum under these three learning settings, respectively.

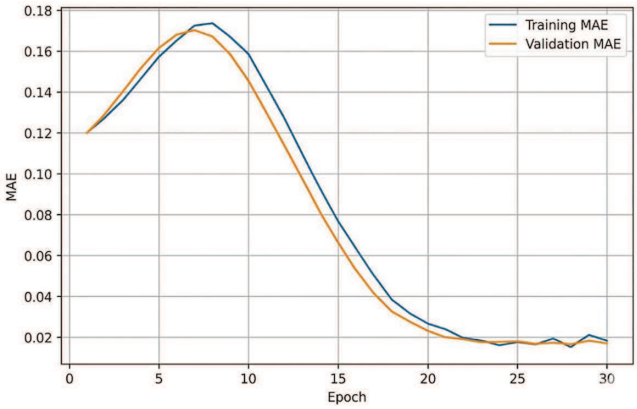
The BiLSTM–8 Heads configuration also showed a competitive performance but was not necessarily better than the proposed four-head model. The relatively small differences between the two configurations indicate that beyond four attention heads, it is not beneficial to these benchmark datasets. One potential reason is that adding extra heads makes it more difficult to learn the representations while having little benefit for modeling the context, especially given the relatively small training sets. Transformer-based architectures also report the need for careful selection of the number of attention heads to keep the model complexity manageable, have good representation and without over-parameterizing the model [32].

5.4.2. Impact of training sample size

The second part of the ablation study examined the effect of varying training sample sizes (15K, 35K, and 60K sampled frames) on the performance of the test variants of the BiLSTM. This is seen in Figure 6, where the number of training samples had only a slight impact on the F1-scores for both benchmark data sets. The BiLSTM–4 Heads configuration had the highest F1-scores in all three training scenarios, and the other attention-based models showed similar performance patterns with slight variations.

The above observations suggest that the proposed attention mechanism is able to effectively learn temporal dependency even with relatively smaller datasets. However, the number of training frames sampled alone did not always lead to an increase in summarization performance, suggesting that attention-head configuration plays a more important role in determining model effectiveness than does the amount of training data. Overall, the results show that the proposed framework shows stable and consistent performance, with the performance being the same for different number of training samples, without significantly reducing the training sample size.

Figure 6
Training accuracy curve of the BiLSTM-Attention model showing stable convergence



5.4.3. Design implications

The design knowledge gained from the ablation experiment can be distilled into the following:

The four-head attention configuration is an effective balance between the temporal feature representation and model complexity.

Reducing the number of attention heads from four is not enough to cause significant performance changes, although there may be some minor differences. There's only a small amount of performance variation when increasing the number of attention heads above four, though some minor changes might happen.

It is not necessarily true that the larger the training set the better the summarization performance becomes when it is greater than 15K.

The proposed framework performs well with various training sample sizes for both benchmark datasets.

5.4.4. Summary

Ablation study gives a better understanding of contribution of individual architectural components to the proposed framework. The BiLSTM–4 Heads configuration consistently outperformed the rest of the evaluated configurations in terms of the F1 score and had a good balance between the complexity of the model and the summarization performance. Moreover, the analysis shows that the attention-head configuration is more effective in improving the performance than just scaling up the amount of training samples, which highlights the effectiveness of the proposed architecture in both of the benchmark sets.

5.5. Final result comparison and discussion

The key novelty of the proposed framework is the seamless combination of BiLSTM-based temporal modeling, multi-head attention, and reconstruction-driven salience estimation in an integrated unsupervised video summarization framework. The proposed approach does not present a fundamentally new architectural element but rather the complementarity of the mechanisms that will enhance the representation of temporal features without increasing the size and computational complexity of the design. The reconstruction error is used as an intrinsic

measure of the importance of the frame and allows generating the summary without any external annotations or reinforcement signals.

As opposed to adversarial or reinforcement learning methods, the proposed approach directly infers the importance of each frame from the reconstruction behavior and thus offers a simple yet reproducible keyframe selection mechanism. The BiLSTM and multi-head attention enable the model to learn the sequential dependence and contextual information, and the reconstruction loss helps select key video segments. As a result, the framework offers an easily understandable alternative to more sophisticated unsupervised summarization solutions.

Based on the experimental evaluation, the proposed BiLSTM–4 Heads model consistently outperforms the other models on the benchmark datasets when trained on 15K dataset and tested on SumMe and TVSum with F1 scores 58.96% and 67.21%, respectively. Furthermore, the proposed integration is also compared with representative unsupervised approaches, demonstrating that the proposed integration gives competitive summarization performance without having complicated model architecture. From these results, we conclude that a proper modeling of the temporal dimension, the use of attention mechanisms, and the reconstruction-based scoring is an effective framework for full unsupervised video summarization.

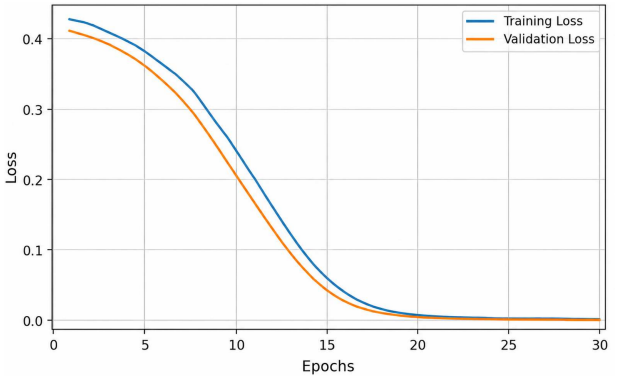
5.5.1. Quantitative results

The training accuracy is shown along with the training behavior of the proposed model in Figure 6. Likewise, in Figure 7, the reconstruction loss during the optimization is gradually decreased, suggesting stable learning behavior during training phase. The results in these observations indicate that the goal of the proposed reconstruction-based approach successfully captures representative temporal features in a video summarization task without supervision.

The experimental results on the two benchmark datasets prove the effectiveness of the proposed BiLSTM–4 Heads framework and indicate that the proposed summarization framework achieves competitive summarization performance. The proposed model achieved the best results in terms of F1 score, 58.96% on SumMe, and 67.21% on TVSum and a good balance between Precision and Recall under the 15K training regime.

Overall, the training and reconstruction-loss curves indicate stable model optimization and support further analysis of the proposed framework's summarization performance.

Figure 7
Loss reduction curve of the BiLSTM-Attention model



5.5.2. Visual representation and variability analysis

The distribution of F1-scores per video on the SumMe dataset is presented graphically in Figure 8, which shows the range of accuracy between the videos. The results show that for most of the analyzed videos, the summarization performance was relatively consistent, with moderate variations seen for videos with dynamic fast cuts from one scene to another or with highly varied visual content.

The differences can be explained by the added degree of complexity in some videos, making it more difficult to determine representative segments. Also, when visually similar frames with semantically different meanings are included in the same videos, the discriminative ability of the reconstruction-based frame importance estimation method may be impaired. The proposed framework, however, shows stable performance on the benchmark dataset, while keeping competitive performance across a variety of video characteristics.

5.6. Cross-dataset generalization

The results clearly demonstrate that the proposed framework performs well in summarization task with both similar and dissimilar datasets. SumMe includes more variable annotation and video categories of scene content, while TVSum includes more of the comparatively structured video categories. Good competitive performance on both datasets indicates the effectiveness of the proposed BiLSTM–Attention framework in capturing the temporal dynamics and capturing the context under different video characteristics.

An exploratory cross-dataset evaluation was also carried out to further explore the transferability of the proposed approach. Retraining the model on TVSum and directly evaluating it on SumMe resulted in a drop of 4.1% in F1-score which could be explained with the more diverse scenes, shorter duration of clips and more subjective annotations in SumMe. When training on SumMe and testing on TVSum, on the other hand, it was found that there was a 3.6% decrease, which suggests that there is a difference in the domains of the user-generated videos and structured videos. While degradation was noted, the modest level of degradation shows that the proposed framework is reasonably cross-dataset robust. We plan to explore lightweight fine-tuning, contrastive pretraining, and domain-adaptive learning to further enhance cross-dataset generalization in future work.

5.7. Comparative performance

Table 5 shows the comparison of the proposed model with the best-performing unsupervised models. A majority of the previous works is based on adversarial training, reinforcement learning, or complicated hierarchical networks. The graphical comparison of the proposed model with state-of-the-art methods is presented in Figure 9. Although our model is a conceptually simpler one, it outperforms other models because of its reconstruction-driven frame importance estimation that is complemented by attention-enhanced sequence modeling.

Our findings indicate that a reconstruction-based model with attention, if trained properly, can achieve better results.

Figure 8
Distribution of F1-scores across SumMe dataset

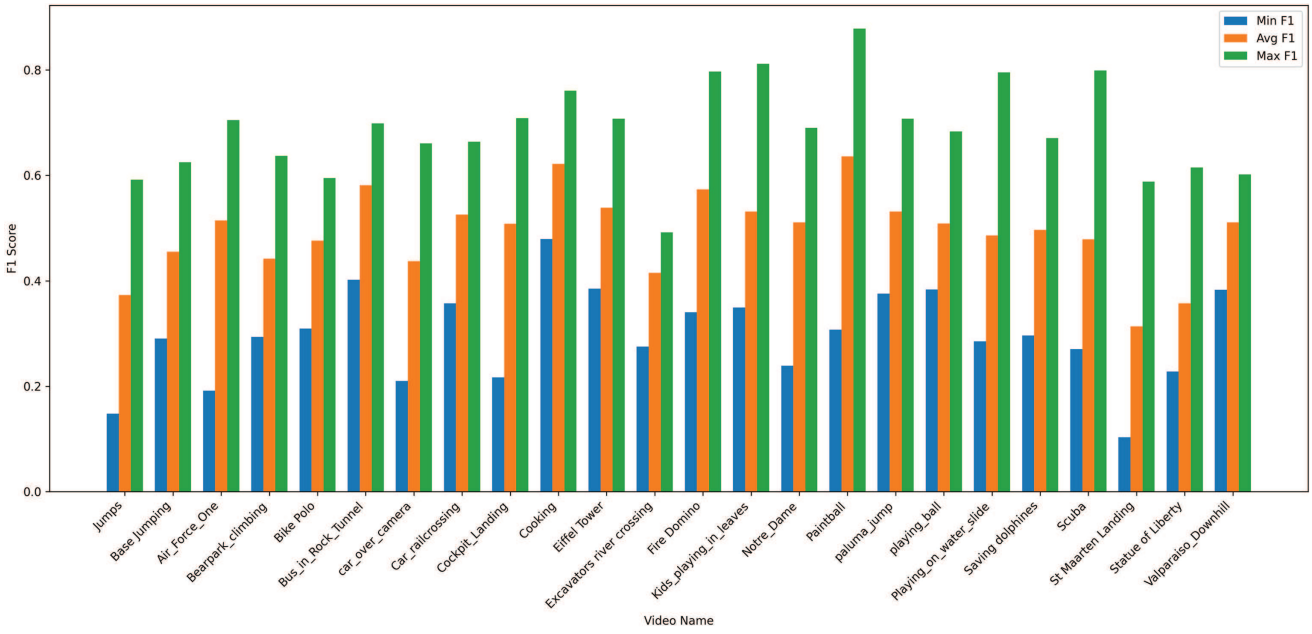


Table 5
Comparison with existing unsupervised video summarization methods

Method	SumMe	TVSum
DR-DSN [41]	41.4	57.6
Cycle-SUM [42]	41.9	57.0
BiLSTM [43]	51.15	60.17
AC-SUM-GAN [44]	50.8	60.6
MAR-Net [45]	49.7	60.4
SUM-GAN [46]	41.7	56.3
BiLSTM-4 Heads (Proposed)	58.96	67.21

5.8. Interpretability and practical implications

One of the critical benefits of the proposed framework is that it is easy to read. We use reconstruction error to estimate the importance of the frame, based on which the key frames are selected, and this is a criterion that is easy to understand without the need to use adversarial training or reinforcement-learning goals. The reconstruction-driven scoring simplifies summarization and maintains the fully unsupervised learning paradigm.

Figure 10 shows example keyframes that were selected by the proposed framework and the example user-annotated keyframes from SumMe and TVSum datasets. The visual comparison shows that the proposed model is able to recognize the main events and

Figure 9
Comparative analysis

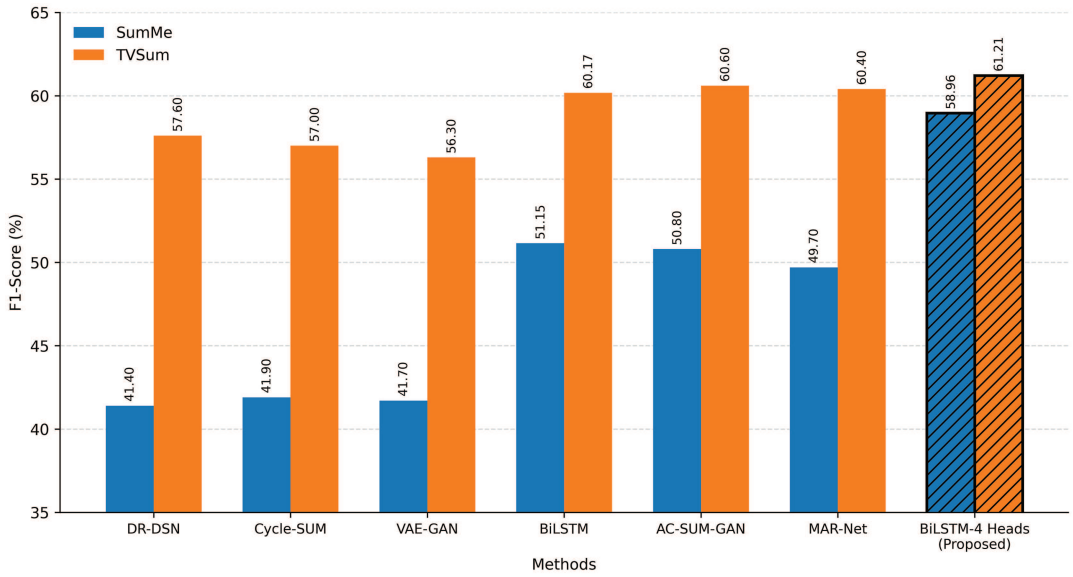


Figure 10
Comparison of top 10 model-selected and user-annotated frames from sample SumMe and TVSum videos



the temporal important segments for each video category while keeping compact summaries of each video category.

Moreover, the proposed architecture integrates ResNet50 feature extraction, BiLSTM temporal modeling, and a moderate four-head attention mechanism in an efficient reconstruction-based model. This design does not require manual annotations in the learning phase and might be applicable to future systems for video summarization that require lightweight unsupervised models.

5.9. Limitations

The proposed framework is successful in both benchmark datasets but there are several limitations. This is because fixed frame sampling might miss out on short-duration events in a video that has fast-changing scenes. Second, the importance score estimation by reconstruction may not give high importance to frames that are semantically similar to other frames but have high visual similarity. Moreover, the existing one is based on visual features from single frames without making use of complementary modalities like audio or text. Lastly, the experimental assessment was tested on benchmark video sets that contain relatively short videos and generalization of the proposed framework to longer and more varied real-world videos is left for future research.

6. Conclusion and Future Scope

In this paper, a compact reconstruction-based unsupervised video summarization framework was proposed, which combines temporal modeling by BiLSTM with an effective keyframe selection method based on four-head attention. The proposed framework uses reconstruction error to estimate frame importance, offering a clear scoring strategy that does not involve manual annotations, adversarial learning, or reinforcement learning. Results from the experimental evaluation on the benchmark SumMe and TVSum datasets showed that proposed BiLSTM–4 Heads configuration led to the best performance with F1-scores of 0.5896 and 0.6721, respectively. Ablation study further showed that a moderate four-head attention configuration could achieve the best balance between contextual representation and model complexity and improve the baseline BiLSTM models modestly and consistently.

Future work will explore adaptive frame sampling techniques, multimodal feature integration with the integration of audio and textual information, and evaluation of larger and more diverse video datasets. In addition, the efforts will be continued on improving cross-dataset generalization and computational complexity in order to enable video summarization applications to be deployed efficiently.

Ethical Statement

This study does not contain any studies with human or animal subjects performed by any of the authors.

Conflicts of Interest

The authors declare that they have no conflicts of interest to this work.

Data Availability Statement

The data that support the findings of this study are openly available in Zenodo at https://zenodo.org/records/4884870 and Github at https://github.com/yalesong/tvsum.

Author Contribution Statement

Manjusha Jagannath Yeola: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Writing – original draft, Writing – review & editing, Visualization. **Sunita S. Barve:** Writing – original draft, Writing – review & editing, Visualization, Supervision, Project administration.

References

[1] Zang, S. S., Yu, H., Song, Y., & Zeng, R. (2023). Unsupervised video summarization using deep non-local video summarization networks. *Neurocomputing*, *519*, 26–35. https://doi.org/10.1016/j.neucom.2022.11.028

[2] Liu, H., Jin, H., Wang, Y., & Yu, H. (2026). Semantic-guided unsupervised video summarization. *arXiv Preprint: 2601.14773*. https://doi.org/10.48550/arXiv.2601.14773

[3] Huang, J., Xin, Y., Qian, J., & Dong, Y. (2025). Multi-temporal granularity concept induction for semantically driven video summarization. *Expert Systems with Applications*, *276*, 127128. https://doi.org/10.1016/j.eswa.2025.127128

[4] Lan, L., Jiang, L., Yu, T., Liu, X., & He, Z. (2026). FullTransNet: Full transformer with local-global attention for video summarization. *IEEE Transactions on Emerging Topics in Computational Intelligence*, *10*(2), 1185–1200. https://doi.org/10.1109/TETCI.2025.3634718

[5] Gupta, J., Rai, D., & Singh, P. (2025). GenSumNet: A genre-specific and context-aware video summarization framework with adaptive thresholding and hybrid temporal modeling. *Expert Systems with Applications*, *279*, 127360. https://doi.org/10.1016/j.eswa.2025.127360

[6] Yu, Q., Wang, Z., Wei, G., & Yu, H. (2026). Deep learning for video summarization: Systematic review, challenges and opportunities. *IEEE/CAA Journal of Automatica Sinica*, *13*(1), 21–42. https://doi.org/10.1109/JAS.2025.125864

[7] Krichen, M., & Mihoub, A. (2025). Long short-term memory networks: A comprehensive survey. *AI*, *6*(9), 215. https://doi.org/10.3390/ai6090215

[8] Srivastava, N., Mansimov, E., & Salakhutdinov, R. (2015). Unsupervised learning of video representations using LSTMs. In *Proceedings of the 32nd International Conference on Machine Learning*, *37*, 843–852. https://doi.org/10.5555/3045118.3045209

[9] Basu, A., Pramanik, R., & Sarkar, R. (2024). Wanet: Weight and attention network for video summarization. *Discover Artificial Intelligence*, *4*(1), 5. https://doi.org/10.1007/s44163-024-00101-y

[10] Otani, M., Song, Y., & Wang, Y. (2022). Video summarization overview. *Foundations and Trends in Computer Graphics and Vision*, *13*(4), 284–355. https://doi.org/10.1561/0600000099

[11] Vora, D., Kadam, P., Mohite, D. D., Kumar, N., Kumar, N., Radhakrishnan, P., & Bhagwat, S. (2025). AI-driven video summarization for optimizing content retrieval and management through deep learning techniques. *Scientific Reports*, *15*(1), 4058. https://doi.org/10.1038/s41598-025-87824-9

[12] Gupta, D., & Sharma, A. (2023). A comprehensive study of automatic video summarization techniques. *Artificial Intelligence Review*, *56*(10), 11473–11633. https://doi.org/10.1007/s10462-023-10429-z

[13] Fu, H., Wang, H., & Yang, J. (2021). Video summarization with a dual attention capsule network. In *2020 25th International Conference on Pattern Recognition*, 446–451. https://doi.org/10.1109/ICPR48806.2021.9412057

[14] Yuan, Y., & Zhang, J. (2023). Unsupervised video summarization via deep reinforcement learning with shot-level semantics. *IEEE Transactions on Circuits and Systems for Video Technology*, *33*(1), 445–456. https://doi.org/10.1109/TCSVT.2022.3197819

[15] Yoon, U. N., Hong, M. D., & Jo, G.-S. (2023). Unsupervised video summarization based on deep reinforcement learning with interpolation. *Sensors*, *23*(7), 3384. https://doi.org/10.3390/s23073384

[16] Li, H., Klabjan, D., & Utke, J. (2025). Unsupervised video summarization via iterative training and simplified GAN. In *Computer Vision – ACCV 2024: 17th Asian Conference on Computer Vision*, 263–279. https://doi.org/10.1007/978-981-96-0966-6_16

[17] Khan, H., Khan, S. U., Ullah, W., & Baik, S. W. (2025). Optimal features driven hybrid attention network for effective video summarization. *Engineering Applications of Artificial Intelligence*, *158*, 111211. https://doi.org/10.1016/j.engappai.2025.111211

[18] Abbasi, M., Hadizadeh, H., & Saeedi, P. (2026). Reinforcement learning for unsupervised video summarization with reward generator training. *IEEE Transactions on Circuits and Systems for Video Technology*, *36*(3), 3041–3054. https://doi.org/10.1109/TCSVT.2025.3623074

[19] Abbasi, M., & Saeedi, P. (2023). Adopting self-supervised learning into unsupervised video summarization through restorative score. In *2023 IEEE International Conference on Image Processing*, 425–429. https://doi.org/10.1109/ICIP49359.2023.10222350

[20] Muhammad, W., Ahmed, I., Ahmad, J., Nawaz, M., Alabdulkreem, E., & Ghadi, Y. (2022). A video summarization framework based on activity attention modeling using deep features for smart campus surveillance system. *PeerJ Computer Science*, *8*, e911. https://doi.org/10.7717/peerj-cs.911

[21] Qaroush, A. M., Hamza, A., Eiss, A., & Hatem, Y. (2026). Multi-video summarization with vision-language models and hybrid optimization. *Knowledge-Based Systems*, *339*, 115566. https://doi.org/10.1016/j.knosys.2026.115566

[22] Marevac, E., Kadušić, E., Živić, N., Buzađija, N., Tabak, E., & Velić, S. (2025). Multimodal video summarization using machine learning: A comprehensive benchmark of feature selection and classifier performance. *Algorithms*, *18*(9), 572. https://doi.org/10.3390/a18090572

[23] Yuan, F., Huang, X., Jiang, H., Jiang, Y., Zuo, Z., Wang, L., . . . , & Peng, Y. (2025). An xLSTM–XGBoost ensemble model for forecasting non-stationary and highly volatile gasoline price. *Computers*, *14*(7), 256. https://doi.org/10.3390/computers14070256

[24] El-Nagar, G., El-Sawy, A., & Rashad, M. (2024). A deep audio-visual model for efficient dynamic video summarization. *Journal of Visual Communication and Image Representation*, *100*, 104130. https://doi.org/10.1016/j.jvcir.2024.104130

[25] Li, H., Ke, Q., Gong, M., & Drummond, T. (2023). Progressive video summarization via multimodal self-supervised learning. In *2023 IEEE/CVF Winter Conference on Applications of Computer Vision*, 5573–5582. https://doi.org/10.1109/WACV56688.2023.00554

[26] Liu, T., Meng, Q., Huang, J.-J., Vlontzos, A., Rueckert, D., & Kainz, B. (2022). Video summarization through reinforcement learning with a 3D spatio-temporal U-Net. *IEEE Transactions on Image Processing*, *31*, 1573–1586. https://doi.org/10.1109/TIP.2022.3143699

[27] El-Soudy, S. R., El-Rashidy, M. A., El-Fishawy, N. A., El-Sayed, A., El-Samie, Abd., E, F., & Selim, G. E. I. (2025). A comprehensive survey of video summarization techniques with emphasis on real-world applications. In *2025 4th International Conference on Electronic Engineering*, 1–8. https://doi.org/10.1109/ICEEM66692.2025.11225028

[28] Simonyan, K., & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*.

[29] Kanafani, H., Ghauri, J. A., Hakimov, S., & Ewerth, R. (2021). Unsupervised video summarization via multi-source

features. In *Proceedings of the 2021 International Conference on Multimedia Retrieval*, 466–470. https://doi.org/10.1145/3460426.3463597

[30] Xu, W., Wang, R., Guo, X., Li, S., Ma, Q., Zhao, Y., . . . , & Yan, J. (2023). MHSCNET: A multimodal hierarchical shot-aware convolutional network for video summarization. In *2023 IEEE International Conference on Acoustics, Speech and Signal Processing*, 1–5. https://doi.org/10.1109/ICASSP49357.2023.10096265

[31] Terbouche, H., Morel, M., Rodriguez, M., & Othmani, A. (2023). Multi-annotation attention model for video summarization. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 3143–3152. https://doi.org/10.1109/CVPRW59228.2023.00316

[32] Han, T., Zhou, Q., Yu, J., Yu, Z., Zhang, J., & Zhao, S. (2024). Effective video summarization by extracting parameter-free motion attention. *ACM Transactions on Multimedia Computing, Communications, and Applications*, *20*(7), 219. https://doi.org/10.1145/3654670

[33] Xie, J., Chen, X., Zhao, S., & Lu, S.-P. (2024). Video summarization via knowledge-aware multimodal deep networks. *Knowledge-Based Systems*, *293*, 111670. https://doi.org/10.1016/j.knosys.2024.111670

[34] Yazdani, H., Bosaghzadeh, A., Ebrahimpour, R., & Dornaika, F. (2025). A computational–cognitive model of audio-visual attention in dynamic environments. *Big Data and Cognitive Computing*, *9*(5), 120. https://doi.org/10.3390/bdcc9050120

[35] Dai, H. L., Zhang, X., & Yu, H. (2023). An attention-based audio-visual fusion method for short video classification. In *Proceedings of the 2023 7th International Conference on Big Data and Internet of Things*, 28–33. https://doi.org/10.1145/3617695.3617728

[36] Gygli, M., Grabner, H., Riemenschneider, H., & van Gool, L. (2014). Creating summaries from user videos. In *Computer Vision – ECCV 2014: 13th European Conference*, 505–520. https://doi.org/10.1007/978-3-319-10584-0_33

[37] Kanafani, H. (2021). SumMe and TVSum datasets (Ver. 1.0) [Data set]. *Zenodo*. https://doi.org/10.5281/zenodo.4884870

[38] Song, Y., Vallmitjana, J., Stent, A., & Jaimes, A. (2015). TVSum: Summarizing web videos using titles. In *2015 IEEE Conference on Computer Vision and Pattern Recognition*, 5179–5187. https://doi.org/10.1109/CVPR.2015.7299154

[39] Wang, X., Li, Y., Wang, H., Huang, L., & Ding, S. (2022). A video summarization model based on deep reinforcement learning with long-term dependency. *Sensors*, *22*(19), 7689. https://doi.org/10.3390/s22197689

[40] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., . . . , & Polosukhin, I. (2017). Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 6000–6010.

[41] Zhang, K., Chao, W.-L., Sha, F., & Grauman, K. (2016). Summary transfer: Exemplar-based subset selection for video summarization. In *2016 IEEE Conference on Computer Vision and Pattern Recognition*, 1059–1067. https://doi.org/10.1109/CVPR.2016.120

[42] Yuan, L., Tay, F. E., Li, P., Zhou, L., & Feng, J. (2019). Cycle-SUM: Cycle-consistent adversarial LSTM networks for unsupervised video summarization. *Proceedings of the AAAI Conference on Artificial Intelligence*, *33*(01), 9143–9150. https://doi.org/10.1609/aaai.v33i01.33019143

[43] Bin, Y., Yang, Y., Shen, F., Xie, N., Shen, H., & Li, X. (2019). Describing video with attention-based bidirectional LSTM. *IEEE Transactions on Cybernetics*, *49*(7), 2631–2641. https://doi.org/10.1109/TCYB.2018.2831447

[44] Apostolidis, E., Adamantidou, E., Metsai, A. I., Mezaris, V., & Patras, I. (2021). AC-SUM-GAN: Connecting actor-critic and generative adversarial networks for unsupervised video summarization. *IEEE Transactions on Circuits and Systems for Video Technology*, *31*(8), 3278–3292. https://doi.org/10.1109/TCSVT.2020.3037883

[45] Zhang, Y., Liu, Y., Kang, W., & Zheng, Y. (2023). MAR-Net: Motion-assisted reconstruction network for unsupervised video summarization. *IEEE Signal Processing Letters*, *30*, 1282–1286. https://doi.org/10.1109/LSP.2023.3313091

[46] Mahasseni, B., Lam, M., & Todorovic, S. (2017). Unsupervised video summarization with adversarial LSTM networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 202–211.

How to Cite: Yeola, M. J., & Barve, S. S. (2026). Unsupervised Video Summarization via BiLSTM-Attention and Reconstruction-Based Scoring. *Journal of Computational and Cognitive Engineering*. https://doi.org/10.47852/bonviewJCCE62028849