

AI-driven federated unsupervised video summarization with ensemble reinforcement learning and explainable caption generation

Ananda Siddhayya Hiremath^{a,b,*}, Ramesh M. Kagalkar^{c,b}

^a BLDEA's V. P. Dr. P. G. Halakatti College of Engineering and Technology, Vijayapura, Karnataka, India

^b Visvesvaraya Technological University, Belagavi, Karnataka, India

^c Nagarjuna College of Engineering and Technology, Bengaluru, Karnataka, India

ARTICLE INFO

Keywords:

Deep learning
Federated learning
Keyframe selection
Transformer models
Video summarization

ABSTRACT

Video summarization is realized as an automated process of reducing long videos into compact representations that effectively describe their content and significance. Video summarization techniques are limited by their high computational resource needs, lack of ability to model long-term dependencies properly, redundancy from frame-to-frame representation, and lack of interpretability. To address these challenges, this paper proposes a novel unsupervised video summarization framework, BiSpiDRL_DOA, that integrates interpolation, hybrid transformer-based temporal modeling, optimized keyframe selection, and explainable caption generation. Initially, video frames are extracted and enhanced through Kolmogorov-Arnold Neural Network (KANN)-based interpolation and multi-feature fusion. A hybrid Vision Transformer-Swin Transformer (ViT_Swin) model is then employed to estimate frame-level importance scores. Keyframes are selected using a Binarized Spiking Deep Reinforcement Learning (BiSpiDRL) model optimized by the Dream Optimization Algorithm (DOA) to reduce redundancy and improve selection accuracy. Furthermore, a BLIP-based caption generation module combined with Grad-CAM produces interpretable textual summaries in Kannada. To support privacy-preserving deployment, Federated Learning (FL) is incorporated for distributed model training without sharing raw video data. Experimental results demonstrate the effectiveness of the proposed framework, achieving a 0.98 success rate, 0.97 sensitivity, 0.96 PPV, and 0.96 F1-score, indicating its suitability for efficient and interpretable real-time video summarization.

1. Introduction

Video is a source of information that is a moving image generated by electronic devices. Social media platforms generate a large volume of videos, so video summarization is the process of converting a long video into a shorter one without losing any essential information from the original video [1].

1.1. Overview of video summarization

With the rapid growth of digital media platforms like YouTube, Facebook, TikTok, and Instagram, video has become a dominant form of entertainment, data dissemination, and communication [2]. Video summarization is a way to quickly access information contained in videos without having to view the entire video [3]. The process of video summarization keeps the key semantics of the original video, so users

can find and interpret the key messages contained within the original video quickly. The process of summarization also produces dynamic summaries that are more engaging to viewers and provide a more comprehensive representation of time-relevant information [4].

1.2. Importance of video summarization

Video summarization technology is an essential tool for rapidly locating, accessing, organizing, and analyzing video content from multiple fields [5]. Video summarization technology is commonly used in entertainment industries, for example, movie trailers, teasers, and reality show trailers [6]. Video summarization technology can be beneficial in many different ways, for instance, conserving time and enabling users to access the most essential details about an event [7]. For example, in sporting events, video summarization can preserve a moment of excellence; in education, it preserves important information

* Corresponding author. BLDEA's V. P. Dr. P. G. Halakatti College of Engineering and Technology, Vijayapura, Karnataka, India.

E-mail addresses: hiremath.a@gmail.com (A.S. Hiremath), rameshvtu10@gmail.com (R.M. Kagalkar).

<https://doi.org/10.1016/j.array.2026.101181>

Received 30 April 2026; Received in revised form 23 August 2026; Accepted 24 August 2026

Available online 28 August 2026

2590-0056/© 2026 The Authors. Published by Elsevier Inc. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

that students can reference later after their class; in the area of surveillance, it can be applied for efficient video monitoring; and in satellite- or marine-related video processing, such as military, police, etc., it also provides an effective means to analyze the content being processed. Video summarization is a growing component within the digital revolution. These applications highlight the importance of video summarization in modern digital ecosystems [8].

1.3. Video summarization using DL

Video summarization has greatly benefited from recent progress in artificial intelligence, Machine Learning (ML), and DL technologies; in particular, DL models allow for the automatic extraction of both spatial and temporal characteristics from a video and thus produce better-quality summaries than other types of summaries [9]. Deep learning (DL) techniques are very effective in handling the analysis of video content produced at high frame rates, as well as for virtual reality videos and short clips on social networking sites [10]. With the rise in popularity of video sharing platforms such as TikTok, Instagram, and YouTube Shorts, deep-learning video summary models can now be used to understand video content instantly [11]. Some DL video summarization models are Convolutional Neural Networks (CNN), Artificial Neural Networks (ANN), and Graph Neural Networks (GNN) but also traditional models like Random Forest (RF) and Support Vector Machine (SVM) models [12].

1.4. Challenges and motivations

While considerable advances have been made in video summarization, the research area still faces large challenges related to computational complexity, modeling long-range temporal relationships, redundancy of selected keyframe, and generalization of video summarization models across diverse video domains. The video summaries created through the most commonly used methods employ the use of supervised learning techniques that depend heavily on large amounts of annotated data in order for these methods to scale. Most current methods struggle with longer videos and cannot provide interpretability to users, thereby limiting their applicability in real-world scenarios.

Existing video summarization approaches typically employ a single feature extractor, a standalone temporal modeling strategy, or supervised keyframe selection techniques. Such approaches often suffer from limited semantic representation capability, poor long-range dependency modeling, redundancy in selected keyframes, lack of interpretability, and privacy concerns.

Current video summarization methods generally use only one feature extractor, a stand-alone temporal modeling method, or supervised key frame selection methods. The methods mentioned above can have a limited ability to represent meaning, a poor ability to model the long-range dependency, redundancy in the selected keyframes, lack of interpretability, and privacy concerns.

Although several video summarization methods have reported promising performance, simpler architectures based on a single feature extraction network, a standalone temporal modeling strategy, or conventional keyframe selection mechanisms remain insufficient for addressing the diverse challenges associated with real-world video summarization. A single feature extractor may capture only a limited range of semantic and spatial characteristics, reducing the ability to represent complex video content comprehensively. Likewise, relying on a single temporal modeling architecture may emphasize either global dependencies or local contextual relationships, while failing to capture both effectively. Conventional supervised keyframe selection methods further require extensive frame-level annotations and often produce redundant summaries that do not adequately represent the complete video content. In addition, most existing approaches provide limited interpretability and do not address privacy concerns arising from centralized data processing. Therefore, a more comprehensive

framework is required to simultaneously improve feature representation, temporal understanding, keyframe diversity, interpretability, and privacy preservation. This motivation forms the basis for the proposed architecture, where each component is incorporated to address a specific limitation that cannot be effectively resolved using simpler standalone alternatives.

To overcome these challenges, this research suggests the development of an unsupervised video summarization framework which can capture both temporal and semantic information of the video content and minimize the redundancy in generating video summary. This framework does not have to be created manually and hence it is more scalable for multiple video formats.

The proposed framework addresses the above limitations through a carefully designed combination of complementary modules. KANN-based interpolation improves temporal continuity by generating smooth intermediate frame representations. Ensemble feature extraction using ConvNeXt, EfficientNetV2, and ShuffleNet captures both rich semantic information and computationally efficient spatial representations. Hybrid ViT-Swin temporal modeling simultaneously learns global temporal dependencies and local contextual relationships that cannot be effectively captured using either transformer individually. For keyframe selection, BSNN reduces computational complexity through binary spiking operations, while DRL enables annotation-free optimization of summary quality. DOA further refines the selected keyframes by minimizing redundancy and maximizing representativeness. BLIP-based caption generation integrated with Grad-CAM enhances interpretability, whereas FL facilitates privacy-preserving deployment in distributed environments. Therefore, the proposed architecture is not a simple aggregation of models but a task-oriented integration in which each component addresses a specific challenge of video summarization.

The framework further incorporates both optimization and the combination of Explanatory Artificial Intelligence (XAI) techniques in order to develop both efficient video summaries and a means for providing interpretive responses by improving the video frame selection and summation of relevant text and to easily adapt this technology to be able to provide services for real-life applications.

1.5. Research contribution

The contribution of the research can be.

- The proposed BiSpiDRL_DOA framework presents a unified unsupervised video summarization architecture that integrates temporal interpolation, ensemble feature extraction, hybrid transformer-based temporal modeling, optimized keyframe selection, explainable caption generation, and federated learning to address redundancy, interpretability, and privacy challenges.
- The novelty of the proposed framework lies in the synergistic integration of complementary modules, where each component addresses a specific limitation of existing video summarization systems, enabling accurate summarization without requiring frame-level annotations.
- A novel BiSpiDRL_DOA-based keyframe selection mechanism is developed by integrating BSNN, DRL, and DOA to improve keyframe representativeness while reducing redundancy and computational complexity.
- Temporal interpolation using KANN is incorporated to enhance temporal continuity and summary smoothness by generating informative intermediate frame representations.
- Ensemble feature extraction and hybrid ViT-Swin transformer-based temporal modeling are employed to effectively capture spatial semantics and long-range temporal dependencies for accurate frame importance estimation.
- Caption generation for the selected keyframes is performed using a BLIP model integrated with Grad-CAM-based explainable artificial

Table 1

Recent work comparison.

Author	Focus	Technique	Advantage	Limitations
Zhang & Liu (2024)	Context-aware video summarization	Bidirectional Graph Matching with contextual correspondence	Captures long-range temporal dependencies and semantic relations effectively	High computational complexity; graph construction overhead
Vo & Vo (2025)	Unsupervised video summarization	Attention-based temporal modeling	Learns temporal dependencies without annotations	Performance degrades for very long videos
Hu et al. (2025)	Long-form video summarization	Meta-prompting with Large Language Models (ViSMaP)	Removes need for manual annotations; strong cross-domain generalization	High computational cost and LLM dependency
Li et al. (2025)	Graph-based video summarization	Spiking Variational Graph (SpiVG) Network	Energy-efficient, captures temporal uncertainty effectively	Complex training; stability issues in spiking networks
Veesam and Satish (2025)	Crime scene video summarization	Multimodal fusion with YOLO-based object detection	Accurate object-aware summarization for surveillance	Domain-specific and limited generalization
Priyanka et al. (2025)	Sports video summarization	Dynamically Hashed Multimodal DL	Efficient feature fusion; reduces redundancy	Requires synchronized multimodal data
Lu et al. (2025)	Event-based video interpolation	Self-supervised event-guided learning	Handles rolling shutter and motion blur effectively	Requires event cameras; limited availability
Ingle & Kim (2025)	Surveillance video synopsis	Panoramic video summarization framework	Efficient for long-duration surveillance videos	Reduced semantic richness; static camera assumption
Zhang et al. (2025)	Video saliency prediction	SETR-Net with SCE and SAT modules	Improves temporal modeling and prediction accuracy	Focused mainly on saliency prediction tasks
Zhang et al. (2025)	Perceptual video compression	SJ-PVC with saliency and JND integration	Reduces bitrate while maintaining visual quality	Primarily optimized for video compression applications
Zhang et al. (2025)	Semi-supervised video object segmentation	Pyramid-structured multi-scale transformer with adaptive fusion	Improves segmentation efficiency and reduces memory usage	Focused on video object segmentation applications
Zhang et al. (2025)	Driver attention prediction	SDFF-Net with dual-branch spatiotemporal fusion	Improves real-time attention prediction with strong spatial-temporal modeling	Designed specifically for driving scenarios
Zhu et al. (2020)	Supervised video summarization	DSNet	Captures temporal consistency and long-range dependencies	Requires supervised training

intelligence, enabling interpretable and semantically meaningful textual summaries.

- A privacy-aware federated learning framework is incorporated to support distributed model training without sharing raw video data, facilitating secure deployment in real-world environments.

The remainder of this paper is organized as follows: Section 2 presents the literature survey, Section 3 describes the proposed framework, Section 4 discusses the experimental results and analysis, and Section 5 concludes the paper with future research directions.

2. Literature survey

In 2024, Y. Zhang and Y. Liu [13] presented the Bgm4Video model, a unique video summarization model that analyzes contextualized relationships across various temporal inputs, based on the graph-matching process. The Bidirectional Graph Transmission (BGT) model analyzes the multi-granularity temporal cues, and a Temporal Encoder (TE) examines both coarse-grained and fine-grained contextual information within videos, forming the two main components. In 2025, B.Q. Vo and V.H. Vo [14] introduced SegSum, an unsupervised model intended for video summarization via the production of video skims. The majority of modern techniques entail training a summarizer to give individual video frames relevance values. These scores are then combined to determine scores for video segments generated through Kernel Temporal Segmentation (KTS). The ViSMaP model was introduced by J. Hu et al. [15] in 2025 and performs summarization through meta-prompting. It leverages LLMs to generate high-quality pseudo-summaries from short video clip descriptions and iteratively refines them through a multi-LLM optimization process. In 2025, W. Li et al. [16] introduced an approach that improves information density while lowering computing cost via the Spiking Variational Graph (SpiVG) Network. It uses a dynamic aggregation graph reasoned that models contextual consistency and semantic coherence across frames in conjunction with a SNN-based keyframe extractor that automatically captures noteworthy temporal occurrences. A Variational Inference Reconstruction A module optimized using Evidence Lower Bound Optimization (ELBO) is used to manage the variability of multi-channel feature fusion, allowing for robust representation learning and enhanced generalization.

In 2025, S.B. Veesam and A.R. Satish [17] suggested the system focus on many camera streams that fuse multiple models and spatial analysis. It combines the detection of a person through an attention-guided Multimodal Fusion Network (AGMFN) and identity matching through the multi-head attention mechanism. 3D CNN and temporal-based networks work together to perform the location analysis of actions and critical frames. DRL-based analysis is also included within the framework for better output productability with enhanced adaptability of the system over the existing system. In 2025, G. Priyanka et al. [18] suggested an application of DL to extract information from multimodal input sources, such as audio from voices and video of cheering crowds as well as each player's facial expressions, using a method they term dynamically hashed multimodal DL (DHMDL). The authors utilized merge sorting as a method of synchronizing the classified video frames based upon scoring systems to allow for parallel processing of the video stream. In a similar manner, Y. Lu et al. [19] discussed the methods of recovering high-quality video frames from event camera data in the presence of Rolling Shutter (RS) distortion and low frame-rate limitations. They describe using event cameras to generate Rolling RS-to-Global Shutter (GS) mappings; inverse GS-to-RS reconstruction, and RS frame warping techniques, thereby allowing for the application of robust mutual supervision of reconstructed video frame data in the absence of any ground truth reference data. In 2025, P.Y. Ingle and Y.G. Kim [20] introduced a panoramic-based video synopsis to solve and analyze the issues of objects for better governance and storage. The system includes multiple cameras sharing a common homograph and constructs a panorama by solving the broad viewpoints with significant deviations, collisions, and overlapping among the images.

In 2025, Y. Zhang et al. [21] introduced a Single Feature Enhancement and Temporal Recurrence Network (SETR-Net) for video saliency prediction. The model combines saliency context enhancement and temporal recurrence mechanisms to improve feature representation and long-range temporal dependency modeling, achieving better prediction accuracy than existing methods. In 2025, Y. Zhang et al. [22] suggested an SJ-PVC for perceptual video compression by integrating DL-based saliency prediction and Just Noticeable Distortion (JND). The method employs adaptive quantization and rate-distortion optimization to reduce visual redundancy, achieving significant bitrate savings while preserving visual quality. In 2025, Y. Zhang et al. [23] suggested a

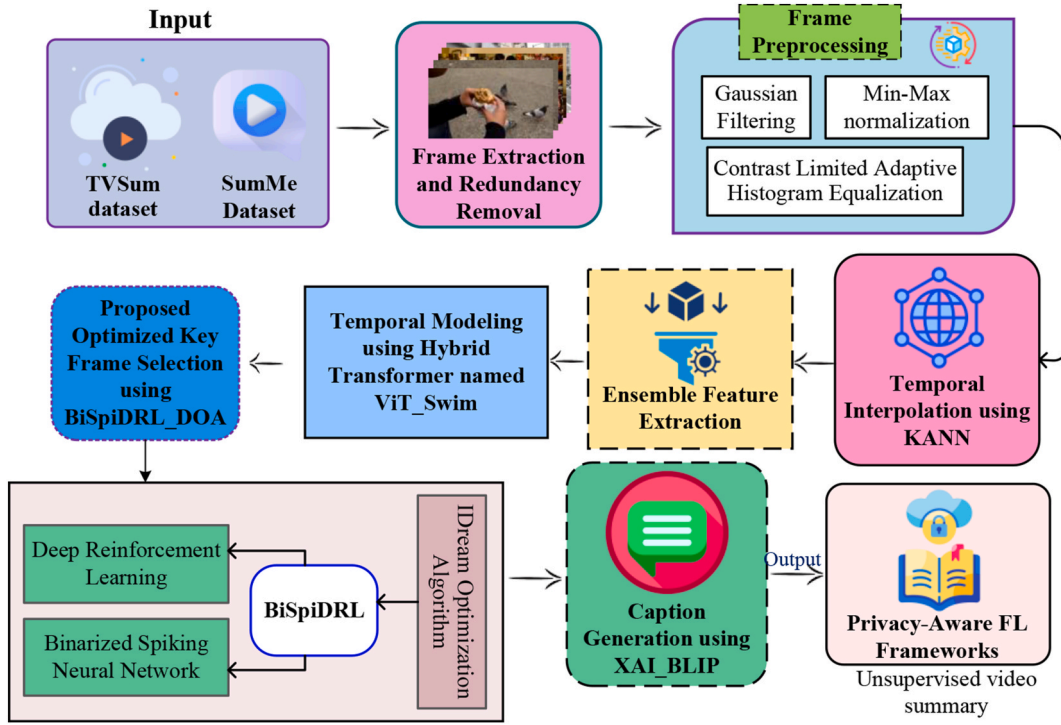


Fig. 1. Schematic representation of overall proposed framework.

pyramid-structured multi-scale transformer for semi-supervised video object segmentation. The framework combines multi-scale feature encoding and adaptive fusion mechanisms to improve object association while reducing GPU memory consumption and maintaining high segmentation efficiency. In 2025, Y. Zhang et al. [24] introduced a Spatiotemporal Dual-branch Feature-guided Fusion Network (SDFF-Net) for driver attention prediction. The model integrates multi-scale spatial feature aggregation and temporal attention transfer mechanisms to effectively capture dynamic scene changes and improve gaze prediction accuracy. In 2020, W. Zhu et al. [25] introduced the Detect-to-Summarize Network (DSNet) for supervised video summarization. The model reformulates video summarization as a temporal interest detection problem rather than a direct regression task. It consists of both anchor-based and anchor-free architectures to identify key video segments effectively. Table 1 presents the recent work comparison.

2.1. Problem statement

Video summarization methods continue to face multiple challenges, such as high computational complexity arising from graph construction overhead and performance degradation when handling very long videos. Several approaches resulted in high computational cost and increased dependency, while others involved complex training with stability issues, particularly in spiking neural networks. Many existing solutions are domain-specific with limited generalization and require synchronized multimodal data, which is not applicable for real-time usage. The insufficiency of proper model for interpolation and the frame selection process leads to video summarization length and makes the process not efficient. Reduced semantic richness and assumptions of static camera conditions further constrain the effectiveness and robustness of current video summarization frameworks.

3. Proposed framework

The study employs a novel framework for the video summarization process, which includes preprocessing, interpolation, feature extraction, temporal modeling, frame selection, and text generation. These stages

collectively illustrate the complete workflow of the proposed method and are visually represented in Fig. 1.

3.1. Video acquisition

The input video is gathered from the TVSum (V_1) and SumMe dataset (V_2) [26]. The dataset V_1 contains 1 to 5 min of 50 YouTube videos, and the size of the dataset is 681.26 MB. The dataset V_2 contains 50 real-world videos with durations ranging from 1 to 6 min, and the size is approximately 24.58 GB, that can be indicated as,

$$V_1 = \{v_1, v_2, v_3, \dots, v_\beta, \dots, v_\alpha\} \quad (1)$$

$$V_2 = \{v_1, v_2, v_3, \dots, v_\gamma, \dots, v_\lambda\} \quad (2)$$

where, α denotes the total count of video in the V_1 dataset, λ is the total count of video in the V_2 dataset, v_β is the β^{th} input video taken for the analysis.

3.2. Frame extraction and redundancy removal

After input collection, the video is converted into frames by using the Uniform Frame Sampling (UFS) [27] under a fixed time interval. In this process the single input video is segmented into multiple shorts, and each short is converted into multiple frames. It can be represented as

$$v = \{s_1, s_2, s_3, \dots, s_n\} \quad (3)$$

where, v is the input video, s_i denotes the i^{th} sort. Each sort consists of samples and is defined as

$$s_i = \{f_{i,1}, f_{i,2}, f_{i,3}, \dots, f_{i,M_n}\} \quad (4)$$

where, $f_{i,1}$ denotes the first frame in the i^{th} sort, and M_n is the total number of frames in the s_i . Once the frames get extracted under a specific time interval, the duplicated frames are eliminated by the similarity matching function. The duplication-removed frames (f) are passed as the input for the preprocessing phase. The process of video-to-frame

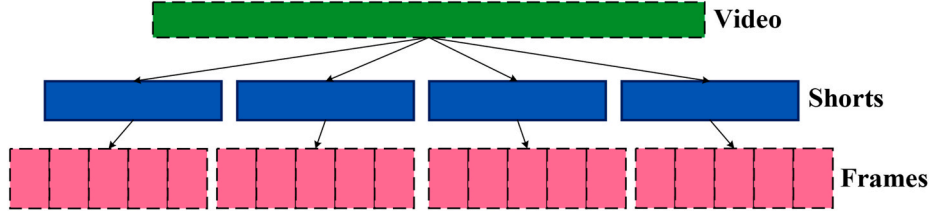


Fig. 2. Process of video-to-frame converting.

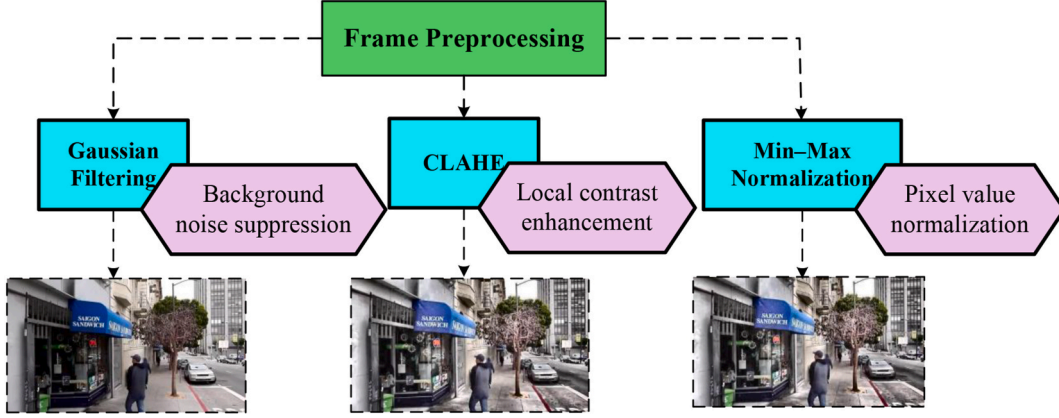


Fig. 3. Process of frame preprocessing.

converting has been presented in Fig. 2.

3.3. Input preprocessing

Preprocessing is the initial stage of computer vision-based DL workflows, which prepares the input frames in a suitable form for further analysis. In this stage, the input frame is treated as an image and undergoes noise removal using Gaussian filtering [28], contrast enhancement through CLAHE [29], and min-max normalization [30] to rescale pixel intensities.

i) Gaussian Filtering

It is a preprocessing technique used to reduce frames background noise and smooth the input by replacing each pixel with the weighted neighboring pixels, and the weights are calculated by using a Gaussian function. It effectively suppresses high-frequency noise while preserving important structural information in the image. The degree of smoothing is controlled by the standard deviation parameter, which determines the width of the Gaussian kernel. It can be represented as

$$g_f(i, j) = \frac{1}{2\pi\sigma_p^2} e^{-\frac{i^2 + j^2}{2\sigma_p^2}} \quad (5)$$

where, i and j denotes the distance on the horizontal and vertical axes, and σ_p represents the standard deviation.

ii) Contrast Limited Adaptive Histogram Equalization (CLAHE)

It is an image-enhancement technique used to improve local contrast in low-light images. It operates on an amplification to avoid excessive noise. The noise-removed input frame is converted to grayscale and divided into small tiles and a histogram of gray-level intensities is computed for each tile. The histograms are clipped using a clip limit T_{clip} , excess counts are redistributed, and a Cumulative Distribution Function (CDF) is used to remap pixel intensities and can be indicated as

$$J_{u,out}(w) = \left| \frac{(U-1)}{N \times P} D_u(w) \right| \quad (6)$$

where, w denotes the input pixel value, $N \times P$ is the tile size, $D_u(w)$ represents the histogram count for gray level, and U is the number of gray levels.

iii) Min-Max normalization

It is a process of rescaling the input frame under pixel intensities to a fixed range, typically $[0, 1]$, that improves model convergence, numerical stability, and robustness. The process shifts each pixel value relative to the minimum intensity and scales it by the overall intensity range of the frame. It ensures that all pixel values contribute proportionally during learning without changing the relative contrast structure.

$$M_{norm} = \frac{M - M_{min}}{M_{max} - M_{min}} \quad (7)$$

where, M denotes the initial value of the frame points, M_{norm} is the frame point normalized value, M_{min} represents the minimum value of frame feature, and M_{max} denotes the frames maximum feature. The overall preprocessing technique and the frame variations have been visualized in Fig. 3. Finally, the preprocessed output (P_{frames}) is passed as the input to the KANN.

3.4. Temporal interpolation using KANN

It [31] is a DL model that contains multivariate functions as compositions of learnable univariate functions. It is chosen for its ability to learn splines to capture complex and nonlinear temporal relationships more accurately. It produces smooth intermediate frames, preserves subtle temporal dynamics, and allows flexible modeling with interpretable spline functions. In this process, the intermediate frames are estimated to eliminate intermediate frame representations, which

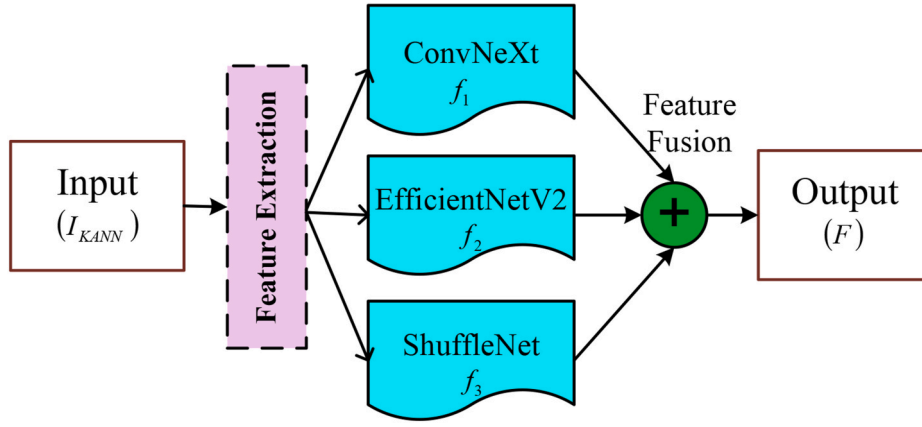


Fig. 4. Process of feature extraction.

results in a smooth motion and provides subtle transitions between frames.

$$g(y_1, y_2, \dots, y_p) = \sum_{p=1}^{2p+1} \phi_r \left(\sum_{q=1}^p \phi_{qr}(P_{frames}) \right) \quad (8)$$

where, $\phi_{qr}(P_{frames})$ denotes the inner univariate transformation of pre-processed frames for intermediate node, ϕ_r is the outer univariate function, $(y_1, y_2, \dots, y_p)$ represents the predicted intermediate frames for interpolation. The KANN output (I_{KANN}), the interpolated frames, is then passed to the feature extraction process.

3.5. Feature extraction

It is the process of converting temporally interpolated frames into high-level numerical representations that capture important semantic and spatial information. ConvNeXt [32], EfficientNetV2 [33], and ShuffleNet [34] are the feature-extracting models that are used to extract the deep visual features.

i) ConvNeXt

It is a CNN-based model designed to capture high-level semantic features and produces rich 1024-dimensional embeddings that represent objects, motion, and context. ConvNeXt was selected over conventional CNN architectures such as VGG and ResNet because of its improved feature representation capability, modernized design, and superior performance in capturing complex visual semantics. The temporally interpolated frames are passed through ConvNeXt to generate dense feature embeddings for each frame. These embeddings serve as input for temporal modeling and keyframe selection. The extracted feature using ConvNeXt can be represented as f_1 .

ii) EfficientNetV2

The process begins by processing the input through the set of MBConv blocks, and each block performs an expanding-compressing process by using deep conv blocks. The Swish activation function helps the network to learn complex relations between pixels and perform smooth nonlinearity. EfficientNetV2 was chosen over traditional CNN models due to its favorable balance between accuracy and computational efficiency, enabling effective feature extraction with reduced training and inference cost. As a result, the hierarchical spatial features are extracted from the input frames, representing semantic, motion, and spatial information for later temporal modeling. The extracted feature using EfficientNetV2 can be represented as f_2 .

iii) ShuffleNet

The inputted frames were processed with the ShuffleNet model, which employs two techniques like group convolution and depthwise separable convolution. These techniques allow for the production of compact frame embeddings while reducing the overall computation needed but still enabling the capture of detailed spatial structure. ShuffleNet was selected because it provides lightweight feature extraction with low computational overhead, making it suitable for efficient video summarization while preserving important spatial details. Residual connections allow for the retention of important information from layer to layer, while the channel shuffle operation permits inter-group communication of all channels within the net. Using this way of compacting the embeddings through model networks, the net produces an encoding for multiscale semantic and spatial representation, that is combined together to provide sufficient information to perform a future temporal representation task. Features obtained from EfficientNetV2 are represented f_3 . The extracted features get fused together for further analysis, and it can be indicated as

$$F = (f_1 + f_2 + f_3) \quad (9)$$

The fused features (F) are subjected to the hybrid transformer model for probability checking among the frames and the ensemble extracting process visualized in Fig. 4.

3.6. Temporal modeling using hybrid transformers

ViT_Swim is a transformer model that integrates the Vision (ViT) [35] transformer for long-range dependencies and the Swim transformer [36] for contextual relationship prediction among features. In this phase of temporal modeling, using a hybrid transformer model resulted in linear prediction among frames that produces frame-level importance probabilities.

i) ViT

It is a DL-based transformer model used to identify the frame-level importance among the probabilities. Among other existing transformer models, ViT has the ability to capture long-range dependencies effectively. It can identify global contextual relationships across frame patches, improving accuracy in complex visual tasks. The input features of frames are split into fixed-size non-overlapping patches, and each patch is flattened and linearly projected into a lower-dimensional embedding with positional encoding. The transformer blocks contain Multi-Head Attention (MHA), normalization, Feed Forward Networks (FFN), and residual connections for the embedding process. It results in linear prediction among frames, producing frame-level importance

probabilities.

$$Attention(R, L, W) = SoftMax\left(\frac{RL^U}{\sqrt{f}}\right) * W \quad (10)$$

where, R is the query, L denotes the key, values are represented by W , f denotes the dimensional between vector R and L . The softmax acts as an activation function, and dense layers with dropout layers are included on the final stage.

ii) Swim

It is a hierarchical version of ViT, and it follows a shifted window mechanism that reduces computational complexity while enabling cross-window connections. It has the ability to capture both local and global contextual information efficiently, making it highly effective for visual tasks. Swin Transformer was selected over conventional transformer architectures because its shifted-window attention mechanism efficiently captures local contextual relationships while maintaining lower computational complexity for high-resolution visual data. The ViT output frames are split into non-overlapping patches, and each of them is represented as a token for local features. These patches are linearly embedded and passed through Swin blocks, while the patch-merging layers perform downsampling, increasing channel dimensions and reducing spatial resolution. The α and δ , combined with λ , β and residual connections extract hierarchical features across the image.

$$\hat{a}_m = \alpha(\beta(a_{m-1})) + a_{m-1} \quad (11)$$

$$a_m = \lambda(\beta(\hat{a}_m)) + \hat{a}_m \quad (12)$$

$$\hat{a}_{m+1} = \delta(\beta(a_m)) + a_m \quad (13)$$

$$a_m + 1 = \lambda(\beta(\hat{a}_{m+1})) + \hat{a}_{m+1} \quad (14)$$

where, λ denotes the MLP, δ is the shifted window-based multi-head self-attention module, β represents the LayerNorm, α is the window-based multi-head self-attention module, a_{m-1} is the input probability, $\hat{a}_m$ and a_m denotes the intermediate and final output block feature. Finally, the frame importance probability output (F_{pro}) is subjected to the key frame selection process for summarization.

3.7. Optimized key frame selection using BiSpiDRL_DOA

The BiSpiDRL network is trained using conventional gradient-based optimization. After training, the candidate keyframes generated by BiSpiDRL are further refined using the Dream Optimization Algorithm (DOA), which performs global search-based keyframe subset optimization to reduce redundancy and improve summary quality.

i) BiSpiDRL

It is an integration of a DL model named Binarized Spiking Neural Network (BSNN) [37] for the initial stage of frame selection and Deep Reinforcement Learning (DRL) [38] for full video summarization. BSNN is also the integration of a Spiking Neural Network and Binarized Neural Network.

3.7.1. BSNN

The model combines Spiking Neural Networks (SNN) and Binarized Neural Networks (BNN) to analyze event-driven spike communication and binary computation. It was selected for its ability to reduce computational cost while maintaining model effectiveness among existing models. Spike trains under the binary weights and activations, enabling faster, energy-efficient, and hardware-friendly implementations. Within the framework, the input is encoded into spike trains,

imitating bioelectric activities observed in biological neurons. Each spike modifies the internal membrane potential of the neuron, and when the threshold is exceeded, an output spike is generated.

$$w_j = \int \sum_k X_{jk} T_k(u) du \quad (15)$$

where, w_j is the membrane potential of neuron j , X_{jk} denotes the synaptic weight, and $T_k(u)$ denotes the is the spike train input. The binarization step then constrains both weights and activations to binary values ($-1, +1$), replacing multiplications with bitwise XNOR operations.

$$y_c = Sign(y) = \begin{cases} -1 & y < 0, \\ +1 & Otherwise \end{cases} \quad (16)$$

where, y denotes the real-valued variables, and y_c is the binarized form. During training, stochastic binarization is used to approximate gradients for back propagation. This introduces randomness based on a probability function, making optimization feasible.

$$\sigma(y) = \min\left(1, \max\left(0, \frac{y+1}{2}\right)\right) \quad (17)$$

where, $\sigma(y)$ denotes the sigmoid function. The BSNN initial stage of frame-selected output FS_{BSNN} is subjected to DRL for the final stage of summarization.

3.7.2. DRL

It is a learning paradigm that integrates reinforcement learning with Deep Neural Networks (DNN) to empower agents to learn optimal decision-making policies directly from high-dimensional state representations via trial-and-error interactions with an environment. This DRL approach is preferred over traditional supervised learning approaches because it efficiently captures long-term temporal dependencies, dynamically adapts to complex state transitions, and removes the need for expensive frame-level annotations. It allows optimal keyframe selection by maximizing long-term rewards while considering summary diversity, representativeness, and computational complexity in an unsupervised learning scenario. Since the framework operates in an unsupervised setting, no manual frame-level annotations or ground-truth keyframe labels are used during training. Instead, the DRL agent learns a keyframe selection policy through reward maximization based on summary diversity, representativeness, temporal coverage, and redundancy reduction. The optimization objective is therefore formulated as a reward-driven learning process rather than supervised classification. The agent assesses its performance by accumulating rewards received from interactions with the environment.

$$S(\tau) = \sum_{u=0}^U s_u \quad (18)$$

$$S(\tau) = \sum_{u=0}^{\infty} \gamma^u s_u \quad (19)$$

where, $S(\tau)$ denotes the cumulative return over trajectory τ , and s_u is the reward received at time step u . To guide decision-making, the agent estimates the expected future rewards associated with states and state-action pairs. These value estimates enable the agent to evaluate long-term consequences of its actions.

$$W(t) = E \left[\sum_{l=0}^{\infty} \gamma^l S_{u+1+l} \mid T_u = t \right] \quad (20)$$

$$R(t, b) = E \left[\sum_{l=0}^{\infty} \gamma^l S_{u+1+l} \mid T_u = t, B_u = b \right] \quad (21)$$

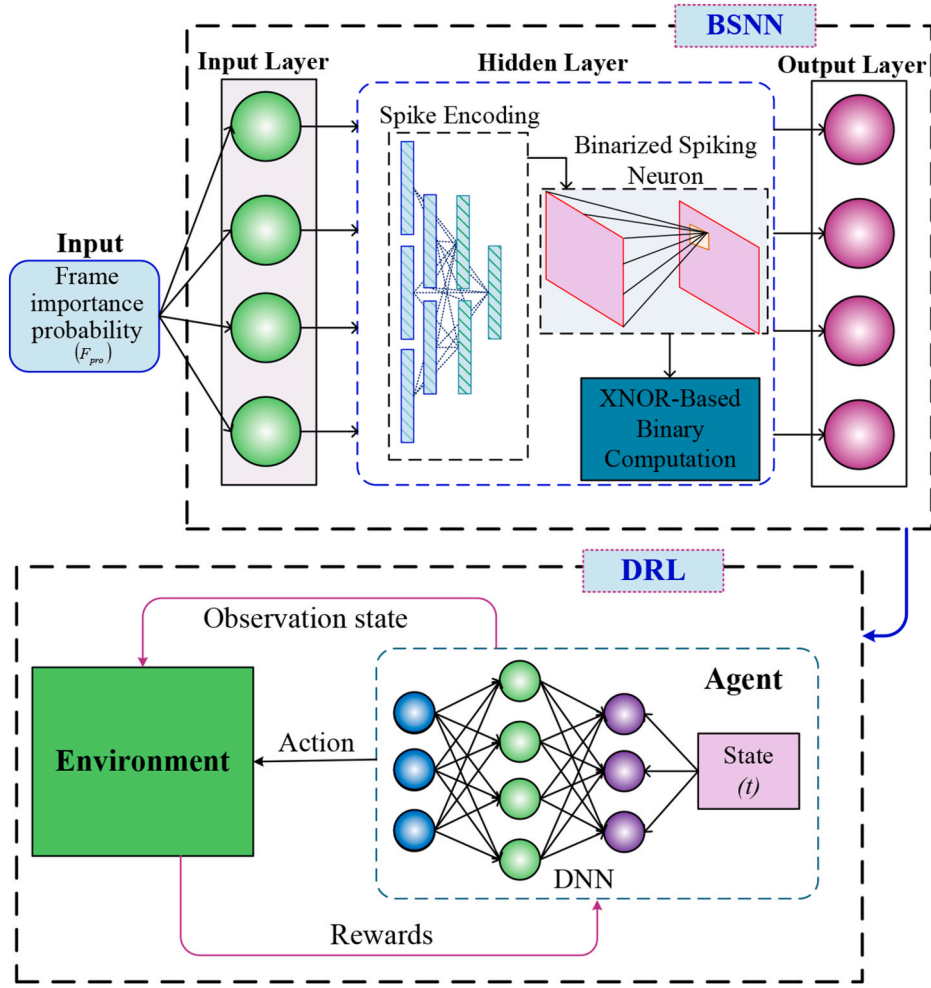


Fig. 5. Architecture of BiSpiDRL.

where, $W(t)$ denotes the state t expected return, and $E[\cdot]$ is the expectation over trajectories. To derive the optimal policy, the agent computes the maximum achievable expected return for each state and state-action pair. This formulation enables action selection that maximizes long-term rewards.

$$W^*(t) = \max_b E[s(t, b) + \gamma W^*(t')] \quad (22)$$

$$R^*(t, b) = E[s(t, b) + \gamma \max_{b'} R^*(t', b')] \quad (23)$$

where, $W^*(t)$ denotes the optimal a state value, $R^*(t, b)$ is the optimal state-action value function, t' represents the next state after b , b' is the next possible action, and $s(t, b)$ denotes the immediate reward. Rather than evaluating actions in isolation, the agent compares each action against the average expected value of the state. This comparison improves learning stability and reduces variance during policy updates.

$$B^\pi(t, b) = R^\pi(t, b) - W^\pi(t) \quad (24)$$

where, $B^\pi(t, b)$ denotes the advantage of state t with action b , $R^\pi(t, b)$ is the state-action value under policy π , and $W^\pi(t)$ is the state-value under π policy. The policy is optimized directly by maximizing the expected return using gradient-based updates. This approach allows learning stochastic policies suitable for complex and continuous action spaces.

$$\nabla_\theta K(\pi_\theta) = F \left[\sum_{u=0}^U \nabla_\theta \log \pi_\theta(b_u | t_u) S(\tau) \right] \quad (25)$$

where, θ denotes the policy parameter, $\pi_\theta(b_u | t_u)$ is the probability action between b_u and t_u , $K(\pi_\theta)$ represents the policy objective function, ∇_θ is the gradient with respect to θ , and $S(\tau)$ denotes the return trajectory τ . The trained BiSpiDRL model generates candidate keyframe importance scores. DOA is subsequently employed to search for an optimal subset of keyframes that minimizes redundancy while maximizing representativeness. Fig. 5 presents the BiSpiDRL model architecture.

ii) DOA

It [39] is a metaheuristic optimization method inspired by the cognitive mechanisms of human dreaming, including memory retention, forgetting, and information reorganization. It is chosen over other algorithms for its ability to adopt forgetting dimensions and dream-sharing mechanisms, which significantly enhance the ability to escape local optima. It has a strong global search capability, adaptive memory-based learning, and high flexibility in handling complex optimization landscapes. The output of the BiSpiDRL is minimized using Mean Squared Error (MSE) as the objective function to preserve essential motion and spatial information while reducing redundancy. To optimize the BiSpiDRL output during video frame selection, ensuring accurate prediction and preserving temporal features is the aim of the optimizer.

3.7.3. Initialization phase

In the initialization phase, DOA generates an initial population randomly within the search space. Let the population matrix be defined as:

$$I = \begin{bmatrix} I_1 \\ I_2 \\ \vdots \\ I_j \\ \vdots \\ I_P \end{bmatrix} = \begin{bmatrix} i_{1,1} & i_{1,2} & \cdots & i_{1,k} & \cdots & y_{1,Dim} \\ i_{2,1} & i_{2,2} & \cdots & i_{2,k} & \cdots & i_{2,Dim} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ i_{j,1} & i_{j,2} & \cdots & \cdots & \cdots & i_{j,Dim} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ i_{P,1} & i_{P,2} & \cdots & i_{P,k} & \cdots & i_{P,Dim} \end{bmatrix} \quad (26)$$

where, P denotes the population size, problem dimension can be represented by Dim , I is the population position matrix, I_1 represents the position of the i^{th} individuals.

3.7.4. Fitness evaluation

It can be indicated as

$$MSE = \frac{1}{B} \sum_{b=1}^B (Op_{BiSpiDRL} - Op_{new})^2 \quad (27)$$

where, B denotes the total number of input frames, Op_{new} is the new optimized output, and $Op_{BiSpiDRL}$ represents the previous stage output.

3.7.5. Exploration phase

In this phase, population is divided into five groups based on the memory capability under the iteration $u = 0$. Different number of dimension contained in the each group $k = L_1, L_2, \dots, L$. On the memory strategy, the each individual resets its initial position to get the best group for the individuals and it be defined as

$$I_j^{u+1} = I_{bests}^u \quad (28)$$

where, I_{bests}^u denotes the best solution that obtained under iteration in the group s . Strategy of forgetting and supplementary, the selected random dimensions $k = L_1, L_2, \dots, L$ position is updated as:

$$i_{j,k}^{u+1} = i_{bests,k}^u + (i_{m,k} + rand \times (i_{v,k} - i_{m,k})) \times \frac{1}{2} \times \left(\cos\left(\pi \times \frac{u + U_{max} - U_e}{U_{max}}\right) + 1 \right) \quad (29)$$

where, u denotes the current iteration, U_{max} is the maximum number of iterations. This process enables a smooth transition from global search to local. In the dream sharing strategy, the individuals may acquire information from randomly selected individuals:

$$y_{j,k}^{u+1} = \begin{cases} y_{n,k}^{u+1} & n \leq j \\ y_{n,k}^u & j < n \leq P \end{cases} \quad (30)$$

where, $n \in [1, P]$, $n \neq j$, and this process increases diversity and helps to escape from the local optima.

3.7.6. Exploitation phase

In this development phase, all individuals follow the global best solutions as from $u = U_e$ to U_{max} , population grouping get removed. The best individual in the entire population (Y_{best}^u) on the memory strategy can be represented as,

$$Y_j^{u+1} = Y_{best}^u \quad (31)$$

$$y_{j,k}^{u+1} = y_{best,k}^u + (y_{m,k} + rand \times (y_{v,k} - y_{m,k})) \times \frac{1}{2} \times \left(\cos\left(\pi \times \frac{u}{U_{max}}\right) + 1 \right) \quad (32)$$

Under strategy of forgetting and supplementation the local exploitation and accelerated convergences get intensified.

3.7.7. Re-evaluating fitness

The fitness is evaluated repeatedly until the system achieves the optimal condition.

3.7.8. Termination

The process terminates once the optimal state is reached. After the termination phase, optimized frame selection is achieved using the novel model, which can be represented as Op_{re-new} . The selected key frames effectively preserve essential temporal and spatial information while minimizing redundancy and error.

3.8. Caption generation using XAI-BLIP

Caption generation is performed using an explainable vision-language framework, which integrates the Bootstrapping Language-Image Pre-training (BLIP) [40] model with an XAI named Grad-CAM [41]. BLIP generates semantically rich and fluent captions for the BiSpiDRL_DOA selected keyframe by jointly learning visual and textual representations. For interpretability, Grad CAM is employed to highlight the discriminative visual regions in each keyframe that most influence the generated captions.

i) BLIP:

The key frames selected by the BiSpiDRL_DOA-based summarization module are forwarded to the BLIP captioning model for semantic description generation. It is a vision-language pre-trained model that jointly learns image-text representations and supports high-quality image captioning. BLIP was selected over conventional image captioning models because its vision-language pretraining enables richer semantic understanding and more contextually accurate caption generation. Given an input keyframe, BLIP encodes the visual content using a vision transformer and generates a corresponding natural language description through an autoregressive text decoder. The captions are generated using beam-search decoding and are constructed to be more fluent and semantically relevant than those generated using greedy decoding because beam-search decoding allows models to explore multiple candidate sequences at each decoding step and select the caption with the highest probability. To eliminate redundancy among generated summaries filtered out any generated captions that communicated similar semantic information through the use of cosine similarity calculations between the bag-of-words representation of the captions. If the cosine similarity value between two captions was greater than a specified threshold, it only retained one of the two captions. This redundancy removal process ensured that the generated captions had a greater level of diversity while preserving the important semantic elements of video summary content. For interpretability analysis, the framework uses an XAI approach called Grad-CAM, which indicates the discriminative visual regions in each keyframe that contribute the most to the generated captions. To support regional language output, the English captions generated by the BLIP model are subsequently translated into Kannada within the module.

ii) Grad CAM

It is an XAI technique used to interpret CNN-based models by identifying the visual regions of an input frame that contribute most to a specific prediction. Grad-CAM was chosen because it provides intuitive visual explanations by highlighting the image regions that most influence the generated captions, thereby improving model interpretability and user trust. It makes use of the gradient information of the final convolution layer, as these layers retain spatial and semantic information that is important for the understanding of textural information. By associating these gradients with feature map activations, XAI provides visual explanations for how certain regions influence the text generation process. The score-based gradients reflect the sensitivity of the output to different spatial regions in the frame. To aggregate this information, global average pooling is applied to the gradients to obtain importance weights for each feature map:

Table 2

Pseudocode for the overall proposed framework.

START
Input data = tvsum_dataset (video), summe_dataset (video)
1. Frame Extraction and Preprocessing
- Extract unique frames from input video
- Apply preprocessing (denoising, contrast enhancement, normalization)
2. Frame Interpolation
- Perform KANN-based interpolation to enrich temporal consistency
3. Feature Extraction and Fusion
- Extract deep features using ConvNeXt, EfficientNetV2, and ShuffleNet
- Fuse multi-model features into a unified feature representation
4. Temporal Importance Estimation
- Apply ViT-Swin temporal model to compute frame importance scores
5. Keyframe Selection using Optimization
- Initialize BiSPiDRL model and DOA optimizer
- Iteratively select candidate frames based on importance scores
- Optimize selection using fitness evaluation (error minimization)
- Obtain optimal keyframes from best DOA solution
6. Caption Generation and Explanation
- Generate captions using BLIP model
- Translate captions to Kannada
- Generate Grad-CAM explanations for interpretability
7. Post-processing
- Remove redundant captions using cosine similarity filtering
- Arrange captions in temporal order
8. Federated Learning Phase
- Distribute local datasets across multiple client nodes
- Train local models at each client without sharing raw video data
- Transmit model parameters to the central aggregation server
- Aggregate local parameters using the FedAvg algorithm
- Update and redistribute the global model across clients
- Repeat the process for multiple communication rounds
Output:
- Optimal keyframes
- Final summarized video captions (Kannada)
- Privacy-preserved global model
END

Table 3

System Configurations.

Components	Specifications
Operating System	Windows 11 (64-bit)
Processor	Intel Core i7 Processor
RAM	16 GB
Pythom	3.14.2
DL Frame	PyTorch

Table 4

Network Configuration settings.

Model	Parameters	Configurations
BSNN	Learning Rate	0.0005
	Activation Function	Sigmoid
	Batch Size	32
	Optimizer	Adam
DRL	Learning Rate	0.0001
	Activation Function	ReLU
	Batch Size	16
	Dropout	0.2
BLIP	Optimizer	Adam
	Learning Rate	0.00005
	Batch Size	8
	Activation Function	GELU
	Pooling Layer	Global Average Pooling

$$\alpha_l^d = \frac{1}{A} \sum_j \sum_k \frac{\partial z_d}{\partial B_{ljk}} \quad (33)$$

where, α_l^d denotes the contribution of l^{th} the feature map to the target output, and B_{ljk} is the activation at spatial location. The weighted feature maps are combined to emphasize the visual regions that most

strongly influence the BLIP-based text generation process. A ReLU function is applied to retain only positive contributions, ensuring that only regions supporting the generated summary are highlighted:

$$M^d = \text{ReLU} \left(\sum_l \alpha_l^d B^l \right) \quad (34)$$

This activation map provides a coarse yet interpretable localization of visually salient content since the generated activation map has a lower spatial resolution than the original input frame; it is upsampled using bilinear interpolation to match the frame dimensions:

$$M_{\text{upsampled}}^d = \text{interpolate}(M_{\text{nor}}^d, \text{size} = (I, X)) \quad (35)$$

where, M_{nor}^d denotes the normalized map, and I, X corresponds to the original frame size. Finally, the framework uses Grad-CAM interpretability analysis and the BLIP-caption generation method for unsupervised video summaries. Redundant captions were discarded for more diverse and relevant generated descriptions, while all remaining unique captions were arranged in chronological order to create the final textual summary. Therefore, this created a final concise, coherent, and informative summary that accurately illustrates the key video summarization elements.

3.9. Privacy-aware FL frameworks

FL is added at the deployment stage for training distributed and private models. The input data is centralized video data spread across numerous client machines, including edge nodes, and mobile systems, with privacy constraints precluding the sharing of raw data. All these parts of the proposed framework, such as preprocessing module, feature extraction module, temporal modelling module, and keyframe selection module, are trained locally on each client rather than sending videos to a central server. The locally updated model parameters are then sent to a central aggregation server where the parameters from all the models are federated averaged to create a global model. This model is distributed back to all clients globally for the purpose of further refining it, so that iterative collaborative learning can take place. The goal of this process is to ensure data privacy, minimize communication overhead and learn from different and large-scale real-world data sources. In the process, several rounds of communication between client and server, and periodic synchronization to ensure model consistency, are repeated. The system can be expanded to accommodate varying needs and conditions. The proposed Privacy-Aware FL Framework was implemented using five distributed client nodes and a central aggregation server. Local model training was performed for five epochs with a batch size of 32 at each client, after which the model parameters were aggregated using the Federated Averaging (FedAvg) algorithm. The global model was updated and redistributed across clients for 50 communication rounds to ensure convergence and model consistency. The overall pseudocode of the proposed framework is summarized in Table 2.

4. Experimental results and discussion

This section proves the efficiency of the proposed model compared to previous ones by demonstrating that the DOA-based keyframe subset optimization improves summary quality compared with other meta-heuristic selection strategies. The proposed model is tested on the TVSum dataset, which contains 50 videos divided into 35 videos (70%) for training, 8 videos (15%) for testing, and 7 videos (15%) for validation, while ensuring that the number of videos is an integer and the ratio of the split is maintained. The captioning process of the chosen keyframes is done by a model through XAI.

The system configuration used in implementing this work is shown in Table 3. The network configuration settings of the DL are shown in Table 4.

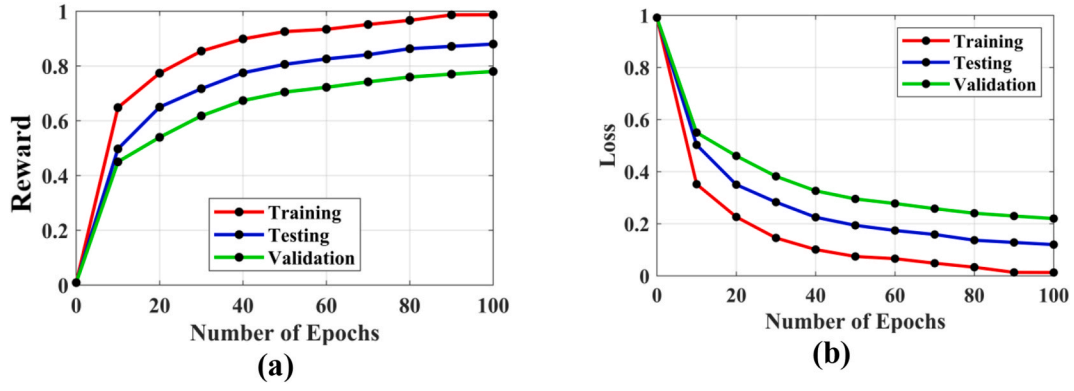


Fig. 6. Training, testing, and validation reward and loss analysis: (a) Reward (b) loss.

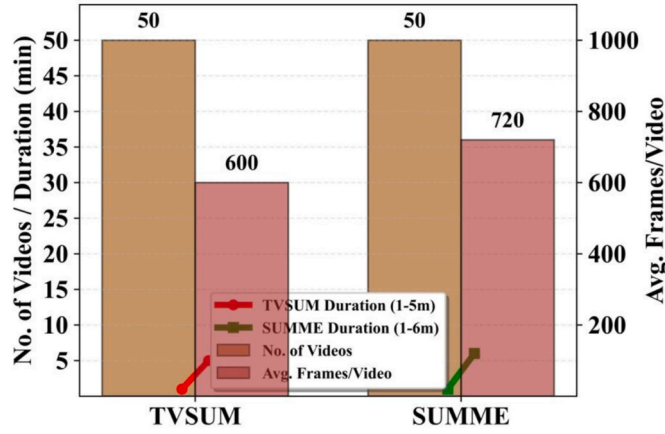


Fig. 7. Dataset analysis.

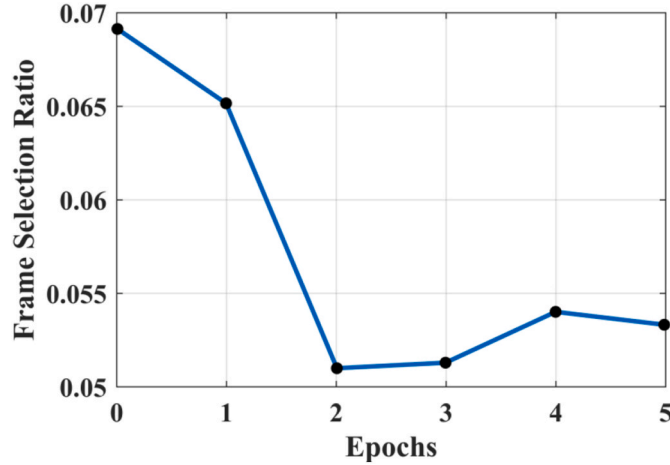


Fig. 8. Ratio of frame selection.

4.1. Performance analysis

The key performance indicators are success rate, PPV, sensitivity, F1-measure, error measures, and computational complexity, which are discussed to show the efficacy of the proposed model. The performance indicators discuss the learning behavior, efficiency, and robustness of the model during training, testing, and validation. The results validate the superior performance of the proposed BiSpiDRL_DOA with XAI_BLP over the existing models. For evaluation purposes, the generated

summaries are compared against reference summaries available in the TVSum and SumMe benchmark datasets. Although the proposed framework is trained in a fully unsupervised manner without utilizing manual frame-level annotations or keyframe labels, benchmark reference summaries available in the TVSum and SumMe datasets are employed exclusively during the evaluation stage to compute Success Rate, PPV, Sensitivity, and F1-score.

Fig. 6 illustrates the training, testing, and validation performance of the proposed framework over 100 epochs. In Fig. 6(a), the reward analysis is presented. The training reward progressively increases from 0.75 at epoch 20 to 0.93 at epoch 60 and reaches its maximum value at epoch 100. Similarly, the testing reward increases to 0.75 at epoch 40, 0.85 at epoch 80, and 0.89 at epoch 100, while the validation reward reaches 0.79 at epoch 100. The continuous increase in reward demonstrates that the DRL agent gradually learns an effective keyframe selection policy through interaction with the environment.

Fig. 6(b) presents the loss analysis over 100 epochs. The training loss decreases from 0.10 at epoch 40 to 0.001 at epoch 100. Likewise, the testing loss decreases from 0.25 at epoch 50 to 0.15 at epoch 100, while the validation loss decreases from 0.30 at epoch 50 to 0.25 at epoch 100. The decreasing loss values indicate stable optimization and convergence of the proposed framework.

Since the proposed framework operates in an unsupervised setting, no manually annotated keyframe labels are utilized during training. Instead, the learning process is driven by reinforcement learning rewards and optimization objectives based on summary diversity, representativeness, temporal coverage, and redundancy reduction. Therefore, the reported curves illustrate model convergence through reward maximization and loss minimization rather than supervised classification accuracy.

Fig. 7 visualizes the dataset distribution analysis datasets in terms of the number of videos, average duration, and average frames extracted per video. The first dataset with durations of 1 to 5 min, for the 50 videos, and 600 frames were extracted per video. The second dataset also contains 50 videos, and durations of 1 to 6 min and 720 frames were extracted per video.

Fig. 8 shows the Frame Selection Ratio (FSR) over the epochs. The video is converted into frames after that. At the end of epochs, the frames are collected. In epoch 1, FSR with 0.065; 0.051 at epoch 2; 0.052 at epoch 3; 0.054 at epoch 4; and 0.053 at the epochs 5.

Fig. 9 shows the existing and proposed model comparison for the modules within it; this highlights the effectiveness of choosing the proposed model over the existing models. In (a), it shows the models used in the interpolation process in the summarization. The existing model CNN has the accuracy of 0.83, ANN has 0.87, and Feedforwarded Neural Networks (FNN) has the accuracy of 0.89; while the proposed KANN model has the accuracy of 0.95 among the existing. This highlights the reason why KANN is chosen over the existing models in the interpolation. In (b), it shows the optimization comparison over the

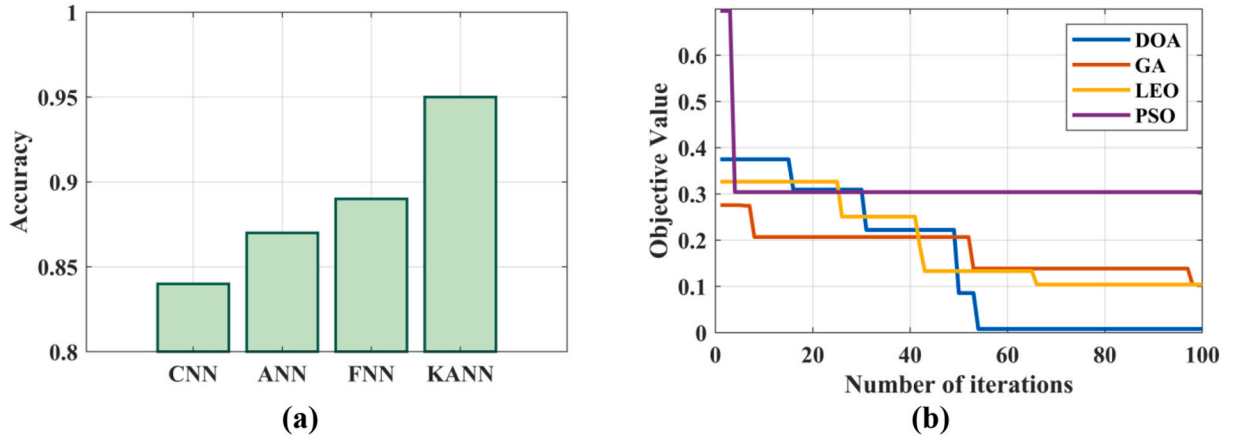


Fig. 9. Performance comparison (a) interpolation models (b) keyframe selection optimization comparison.

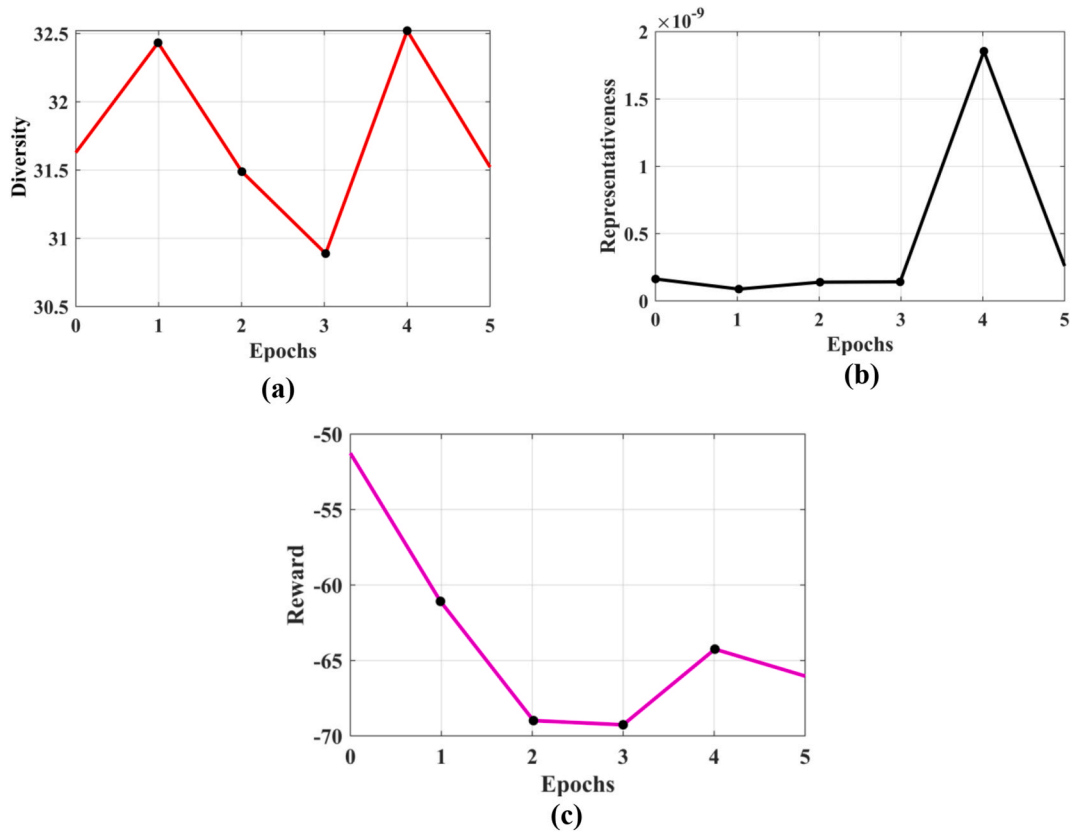


Fig. 10. DRL based Analysis over epochs (a) Diversity Reward (b) Representativeness Reward (c) Total Rewards.

iterations. The existing optimization model GA has the loss value of 0.11, LEO has the loss value of 0.1, and PSO has the value of 0.3; while the proposed model DOA has the lowest loss value of 0.01. It highlights that the proposed model effectiveness has been increased by using the DOA optimization as an optimizer.

Fig. 10 shows the DRL-based analysis in the video summarization process, illustrating overall changes in the reinforcement function over five epochs. In (a) it shows the diversity rewards: 32.4 at epoch 1, 31.5 at epoch 2, 30.8 at epoch 3, 32.5 at epoch 4, and 31.5 at epoch 5. In (b) it shows the representativeness rewards over the epochs: 0.2×10^{-9} at epoch 0, 0.1×10^{-9} at epoch 1, 0.2×10^{-9} at epoch 2, approximately the same at epoch 3, 1.9×10^{-9} at epoch 4, and 0.3×10^{-9} at epoch 5. In (c) it shows the total rewards: -51 at epoch 0, -62 at epoch 1, -69 at epochs 2 and 3, -64 at epoch 4, and -66 at epoch 5.

Fig. 11 shows the performance comparison of existing models and the proposed BiSpiDRL_DOA method for video summarization. In sub-figure (a), the success rate of BNN is 0.78, GNN is 0.81, and BiLSTM is 0.84, whereas the proposed model achieves a higher success rate of 0.96. The PPV value for BNN is 0.79, GNN is 0.79, and BiLSTM is 0.83, while BiSpiDRL_DOA attains 0.94. In sensitivity analysis, the values of BNN, GNN, and BiLSTM are 0.76, 0.80, and 0.81, which are significantly lower than the proposed model value of 0.94. The F1 measures are 0.80 for BNN, 0.83 for GNN, and 0.84 for BiLSTM, while 0.95 for the proposed approach. In (b), it shows the error analysis among models like BNN, which has an MSE of 0.04 and an MAE of 0.03; GNN, which has an MSE of 0.03 and an MAE of 0.03; and BiLSTM, which has an MSE of 0.028 and an MAE of 0.02, while the proposed model overcomes the existing models' performance by achieving an MSE of 0.01 and an MAE of 0.01.

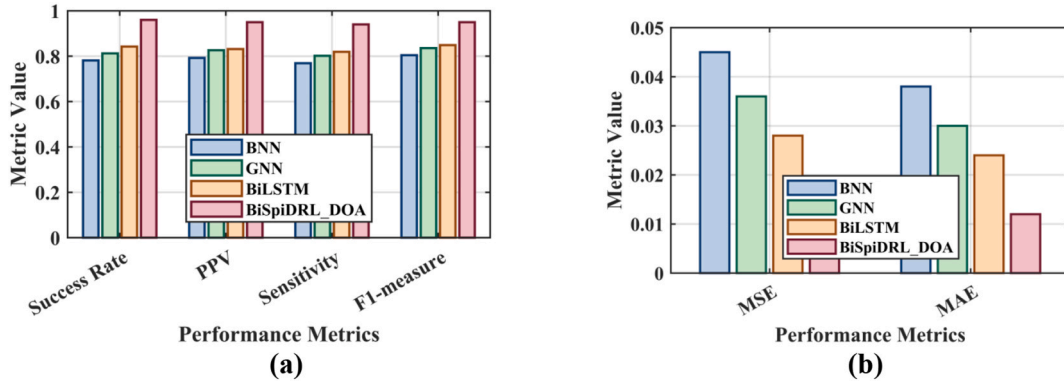


Fig. 11. Existing and proposed model comparison Video summarization:(a) Performance Metrics (b) Error Metrics Analysis.

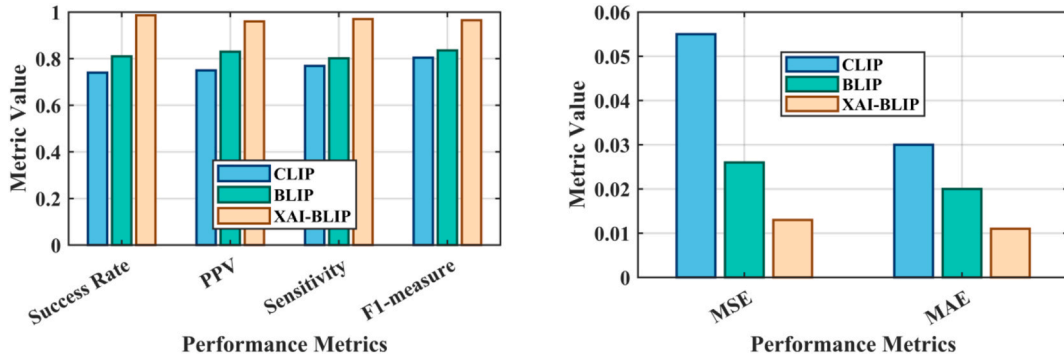


Fig. 12. Comparison for Caption generation Models: (a) Performance Metrics (b) Error Metrics Analysis.

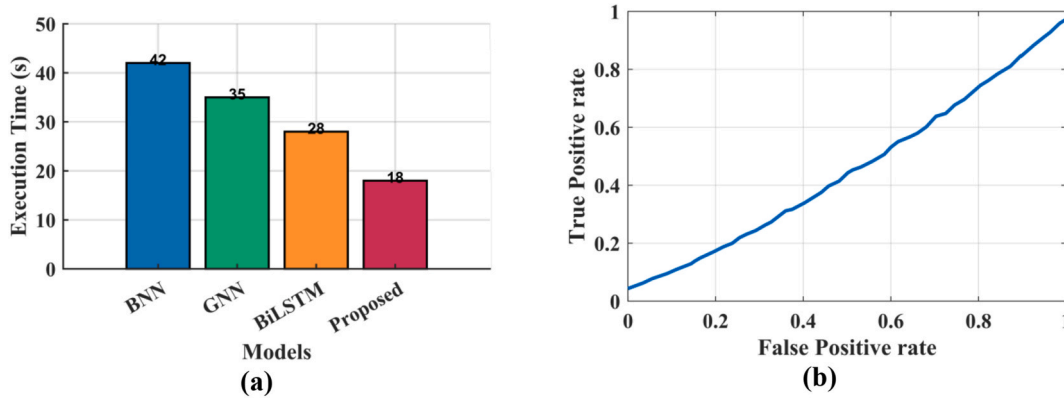


Fig. 13. Performance analysis (a) computational complexity analysis (b) ROC analysis.

Fig. 12 illustrates the performance comparison of caption generation models integrated within the full video summarization framework. Since caption generation operates as a component of the complete BiSpiDRL-based video summarization model, its performance directly reflects the effectiveness of semantic understanding in the summarized video output. In subfigure (a), the Contrastive Language–Image Pre-training (CLIP) model achieves a success rate of 0.74, PPV of 0.75, sensitivity of 0.7689, and an F1-measure of 0.8043, while BLIP improves these values to 0.81, 0.83, 0.8017, and 0.8356, respectively. The proposed XAI-BLIP model performs much better than both, with a success rate of 0.9863, PPV of 0.96, sensitivity of 0.97, and F1-measure of 0.965. In subfigure (b), the error measures again confirm the effectiveness of the proposed method. CLIP has higher errors with an MSE of 0.055 and MAE of 0.030, whereas BLIP decreases the error to an MSE of 0.026 and MAE of 0.020. The XAI-BLIP model again has the lowest error measures with

MSE of 0.013 and MAE of 0.011, indicating better captioning accuracy and semantic correctness.

Fig. 13 illustrates the performance analysis of the model. In (a), it compares the computational complexity analysis models. The existing model BNN has an execution time of 42 s, GNN has 35 s, and BiLSTM model has 28 s, whereas the proposed model has the lowest execution time of 18 s. In (b), it illustrates the ROC analysis. The proposed model has a steadily increasing trend, showing that the model has a better discrimination power. When the true positive rate is increased from 0.2, the ROC curve is close to the upper-left corner of the ideal region, showing that the model has the capability to classify the relevant and non-relevant video segments.

Table 5 presents the visualization outcome of the TVSum dataset and Table 6 presents the outcome of SumMe dataset.

Table 7 shows a comparison of recent video summarization

Table 5
Visualization outcome of TvSum dataset.

Keyframe with Grad-CAM Visualization		
Generated English Caption	A cycling event in a city setting, featuring participants riding along the roadway while spectators observe the competition from the sidelines	A street-side dining scene where a person purchases a burger from a shop and enjoys the meal outdoors, accompanied by numerous sparrows gathered nearby
Kannada Translation	ನಗರದ ಹಿನ್ನೆಲೆಯಲ್ಲಿ, ನಡೆಯುವ ಸೈಕ್ಲಿಂಗ್ ಕಾರ್ಯಕ್ರಮ, ಇದರಲ್ಲಿ ಭಾಗವಹಿಸುವವರು ರಸ್ತೆಯ ಉದ್ದಕ್ಕೂ ಸವಾರಿ ಮಾಡುತ್ತಿದ್ದರೆ, ಪ್ರೇಕ್ಷಕರು ಸ್ಪರ್ಧೆಯನ್ನು ಪಕ್ಕದಿಂದ ವೀಕ್ಷಿಸುತ್ತಿದ್ದಾರೆ	ಬೀದಿ ಬದಿಯ ಊಟದ ದೃಶ್ಯ, ಅಲ್ಲಿ ಒಬ್ಬ ವ್ಯಕ್ತಿ ಅಂಗಡಿಯಿಂದ ಬರ್ಗರ್ ಖರೀದಿಸಿ ಹೊರಗೆ ಊಟ ಮಾಡುವುದನ್ನು ಆನಂದಿಸುತ್ತಾನೆ, ಜೊತೆಗೆ ಹತ್ತಿರದಲ್ಲಿ ನೆರೆದಿರುವ ಹಲವಾರು ಗುಬ್ಬಿಚ್ಚಿಗಳು ಸಹ ಇರುತ್ತವೆ

Table 6
Visualization outcome of SumMe dataset.

Keyframe with Grad-CAM Visualization		
Generated English Caption	A bicycle training session where riders practice cycling skills and stunt maneuvers under the guidance of a coach providing instructions in an indoor arena	Daily activities of young adults, including listening to music, practicing yoga, walking along the street, and capturing photographs using a camera
Kannada Translation	ಒಳಾಂಗಣ ಕ್ರೀಡಾಂಗಣದಲ್ಲಿ ಸೂಚನೆಗಳನ್ನು ನೀಡುವ ತರಬೇತುದಾರರ ಮಾರ್ಗದರ್ಶನದಲ್ಲಿ ಸವಾರರು ಸೈಕ್ಲಿಂಗ್ ಕೌಶಲ್ಯ ಮತ್ತು ಸಾಹಸ ತಂತ್ರಗಳನ್ನು ಅಭ್ಯಾಸ ಮಾಡುವ ಬೈಸಿಕಲ್ ತರಬೇತಿ ಅವಧಿ	ಸಂಗೀತ ಕೇಳುವುದು, ಯೋಗಾಭ್ಯಾಸ ಮಾಡುವುದು, ಬೀದಿಯಲ್ಲಿ ನಡೆಯುವುದು ಮತ್ತು ಕ್ಯಾಮೆರಾ ಬಳಸಿ ಛಾಯಾಚಿತ್ರಗಳನ್ನು ಸೆರೆಹಿಡಿಯುವುದು ಸೇರಿದಂತೆ ಯುವ ವಯಸ್ಕರ ದೈನಂದಿನ ಚಟುವಟಿಕೆಗಳು

techniques on the most popular video benchmark datasets in their respective experimental protocols. A report is generated for each old method performed, with the dataset it was originally validated with, which allows them to be fairly compared with different evaluation settings. The proposed BiSpiDRL_DOA + XAI_BLIP model is tested on a combined TVSum + SumMe data and attains Success Rate of 0.96, PPV of 0.95, Sensitivity of 0.95, and F1-score of 0.95 which is better compared to all the compared methods. This shows great effectiveness of the proposed scheme in the management of multiple video characteristics and in the quality of the selection of the key frames.

To compare the most recent video summarization models to date, we have conducted experiments on two video summarization benchmark sets (TVSum and SumMe) using the same experimental protocol, as shown in Table 8. Evaluations are conducted in the same evaluation environment, making for an equitable comparison of the methods. The results demonstrate that the proposed BiSpiDRL_DOA framework consistently achieves better results compared with the current

approaches on both the datasets. It gets better values of Success Rate, PPV, Sensitivity and F1-score compared with recent models, and it has better quality of keyframe selection and better temporal representation learning.

Table 7
Dataset-wise comparison of summarization methods.

Methods	Dataset	Success Rate	PPV	Sensitivity	F1-score
SegSum	TVSum	0.72	0.69	0.71	0.71
SpiVG	ActivityNet	0.73	0.71	0.69	0.69
AGMFN	SumMe	0.78	0.76	0.	0.75
DHMDL	YouTube	0.84	0.83	0.82	0.81
Proposed (BiSpiDRL_DOA + XAI_BLIP)	Highlights TVSum + SumMe	0.96	0.95	0.95	0.95

Table 8

Comparison of existing and proposed video summarization methods on datasets.

Methods	TVSum Dataset				SumMe Dataset			
	Success rate	PPV	Sensitivity	F1-score	Success rate	PPV	Sensitivity	F1-score
SegSum (2025)	0.84	0.83	0.82	0.83	0.81	0.80	0.79	0.80
SpiVG (2025)	0.86	0.85	0.84	0.85	0.83	0.82	0.81	0.83
AGMFN (2025)	0.88	0.87	0.86	0.87	0.85	0.84	0.83	0.85
DHMDL (2025)	0.87	0.86	0.85	0.86	0.83	0.82	0.81	0.83
SETR-Net (2025)	0.89	0.88	0.87	0.88	0.86	0.85	0.84	0.86
SJ-PVC (2025)	0.88	0.87	0.86	0.87	0.85	0.84	0.83	0.85
SDFF-Net (2025)	0.90	0.89	0.88	0.89	0.87	0.86	0.85	0.87
Proposed (BiSpiDRL_DOA + XAI_BLIP)	0.96	0.95	0.94	0.95	0.95	0.94	0.93	0.95

Table 9

Federated learning performance and privacy analysis.

Method	Success Rate	PPV	Sensitivity	F1-Score	Privacy Level	Raw Data Sharing
Centralized Learning	0.87	0.85	0.86	0.88	Low	Yes
Conventional FL (FedAvg)	0.90	0.91	0.87	0.84	High	No
Proposed Privacy-Aware FL Framework	0.96	0.95	0.95	0.95	Very High	No

Table 10

Ablation study.

Model Configuration	Success Rate	PPV	Sensitivity	F1-measure
Full Model (BiSpiDRL_DOA + XAI_BLIP)	0.98	0.96	0.97	0.96
Without BLIP (No Captioning Module)	0.93	0.91	0.90	0.91
Without BiSpiDRL (Replaced with standard RL)	0.90	0.88	0.87	0.88
Without ViT_Swin (Replaced with simple temporal model)	0.86	0.84	0.83	0.84
Without KANN (No interpolation module)	0.80	0.78	0.76	0.77
Ensemble Feature Extraction Only (Base Model)	0.54	0.52	0.51	0.52

Table 9 presents the performance and privacy analysis of the Federated Learning framework. The results demonstrate that the proposed Privacy-Aware FL framework achieves performance comparable to centralized learning, with a Success Rate of 0.96, PPV of 0.95, Sensitivity of 0.95, and F1-score of 0.95, while ensuring enhanced privacy by eliminating raw data sharing between clients and the central server.

The ablation study as given in Table 10 is a controlled analysis to assess the contribution of each module in the proposed framework. To examine the individual contribution to its overall performance, each component (KANN, ViT_Swin, BiSpiDRL and BLIP) is systematically removed. The outcomes are clearly showing the consistent reduction of the performance measures if any module is removed, which shows the effectiveness and necessity of each module.

4.2. Discussion

The framework illustrates an efficient and robust framework for unsupervised video summarization by incorporating in the pipeline. Interpolation and frame selection in the BiSpiDRL_DOA framework preserve the temporal and semantic information and minimize the duplication in the video content. The Hybrid Transformer Network architecture used by the model allows real-time estimation of frame

importance through the capture of both local and global context dependencies. Together with the adaptive optimization process, spiking computations, and weighing through DRL, models consistently outperform the previous ones. The use of XAI in interpreting the generated video summarization increases the chances of the end-user understanding the intended semantics of the video, which paves the way for using the technique in real world applications like surveillance video applications, content summarization for the media and intelligent video content analysis. It is worth to note that the proposed framework is tested with benchmark datasets (Kaggle based offline video streams), and the experiments are therefore performed for controlled, pre-recorded, video scenarios rather than for real-time video acquisition scenarios. Moreover, the framework can be expanded to include an FL to enable distributed and privacy-preserving model training on multiple devices, further increasing its applicability to real-world, data-intensive environments.

5. Conclusion

This study proposes a novel video summarization system that integrates interpolation, deep feature fusion, temporal modeling, optimal keyframe selection, and text captioning. After video preprocessing, the video frames are merged through interpolation to generate intermediate video frames. Based on the intermediate frames, a semantic representation of the video content is obtained through an ensemble of feature extraction networks. The fused features are modeled in the temporal domain using a hybrid ViT_Swin transformer to predict the importance probabilities of the frames. Keyframe selection was done, and captioning was also integrated into the system. The summarization achieves a high PPV of 0.95 with low error values, an MSE of 0.015, and an MAE of 0.012. Additionally, the caption generation module using XAI_BLIP attained a PPV of 0.96 while maintaining very low errors of an MSE of 0.013 and an MAE of 0.011, highlighting its strong descriptive and contextual understanding capability.

5.1. Research limitations and future works

The limitation of the current research work is that it uses benchmark datasets with offline video streams rather than addressing the issues and difficulties of capturing live video streams and real-time data acquisition. In addition to that, the research work has not assessed the method on hardware resource-limited platforms, and the deployment efficacy on edge devices has not been explored. The future directions include extending the model to handle live videos through online inference mechanisms, incorporating audio and multi-modal inputs for more elaborately phrased captions, and further optimizing the model for deployment on hardware-friendly setups for streaming applications.

CRedit authorship contribution statement

Ananda Siddhayya Hiremath: Writing – original draft. Ramesh M. Kagalkar: Supervision.

Funding

Not applicable.

Declaration of competing interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The datasets used in this study, namely TVSum and SumMe, are publicly available benchmark datasets. TVSum is available at <https://github.com/yalesong/tvsum> and SumMe at <https://gyglim.github.io/me/vsum/index.html>.

References

- [1] Yu Q, Yu H, Sun Y, Ding D, Jian M. Unsupervised video summarization based on the diffusion model of feature fusion. *IEEE Trans Comput Social Syst* 2024;11(5): 6010–21.
- [2] Negi A, Kumar K, Saini P. Object of interest and unsupervised learning-based framework for an effective video summarization using deep learning. *IETE J Res* 2024;70(5):5019–30.
- [3] Khan H, Hussain T, Khan SU, Khan ZA, Baik SW. Deep multi-scale pyramidal features network for supervised video summarization. *Expert Syst Appl* 2024;237: 121288.
- [4] Wu G, Song S, Wang X, Zhang J. Reconstructive network under contrastive graph rewards for video summarization. *Expert Syst Appl* 2024;250:123860.
- [5] Kashid S, Awasthi LK, Berwal K, Saini P. STVS: Spatio-temporal feature fusion for video summarization. *IEEE Multimedia* 2024.
- [6] Huang J, Xin Y, Qian J, Dong Y. Multi-Temporal Granularity Concept Induction for semantically driven video summarization. *Expert Syst Appl* 2025;276:127128.
- [7] Benoughidene A, Titouna F, Boughida A. Static video summarization based on genetic algorithm and deep learning approach. *Multimed Tool Appl* 2025;84(13): 12487–512.
- [8] Sharma S, Khunteta A, Goyal D. DeepK-Means: a fusion of DCNN and K-Means clustering for video summarization. *Int J Algorithm Comput Math* 2025;11(1):1.
- [9] Wu L, Ye K, Gao H, Jiang H, Zhang H. Video frame interpolation based on invertible neural network for internet of things. *Digital Communications and Networks* 2025.
- [10] Sartinis EG, Psarakis EZ, Kosmopoulos DI. A 3D wrist motion-based sign language video summarization technique. *Pattern Recognit Lett* 2025;189:23–30.
- [11] Li Q, Zhan Z, Li Y, Bhanu B. Spatial-temporal multi-scale interaction for few-shot video summarization. *Eng Appl Artif Intell* 2025;142:109883.
- [12] Islam MM, Kakouros S, Heikkilä J, Oussalah M. Towards an automated multimodal approach for video summarization: building a Bridge between text, audio and facial cue-based summarization. *arXiv preprint arXiv:2506.23714* 2025.
- [13] Zhang Y, Liu Y. Contextual correspondence matters: bidirectional graph matching for video summarization. In: *European conference on computer vision*. Cham: Springer Nature Switzerland; 2024, September. p. 300–17.
- [14] Vo BQ, Vo VH. Integrate the temporal scheme for unsupervised video summarization via attention mechanism. *IEEE Access* 2025.
- [15] Hu J, Korkinof D, Gong S, Beguerisse-Diaz M. ViSMaP: unsupervised hour-long video summarisation by meta-prompting. *arXiv preprint arXiv:2504.15921* 2025.
- [16] Li W, Han W, Deng LJ, Xiong R, Fan X. Spiking variational graph representation inference for video summarization. *IEEE Trans Image Process* 2025.
- [17] Veeram SB, Satish AR. Design of an integrated model for video summarization using multimodal fusion and YOLO for crime scene analysis. *IEEE Access* 2025.
- [18] Priyanka G, Senthil Kumar J, Prasha Meena M. DHMDL: dynamically hashed multimodal deep learning framework for racket video summarization using audio and visual markers. *Appl Artif Intell* 2025;39(1):2462382.
- [19] Lu Y, Liang G, Shen Y, Wang L. Self-supervised learning of event-guided video frame interpolation for rolling shutter frames. *IEEE Trans Visual Comput Graph* 2025.
- [20] Ingle PY, Kim YG. Panoramic Video synopsis on constrained devices for security surveillance. *Systems* 2025;13(2):110.
- [21] Zhang Y, Xiao Y, Zhang Y, Zhang T. Video saliency prediction via single feature enhancement and temporal recurrence. *Eng Appl Artif Intell* 2025;160:111840.
- [22] Zhang Y, Zhang T, Wang S, Yu P. An efficient perceptual video compression scheme based on deep learning-assisted video saliency and just noticeable distortion. *Eng Appl Artif Intell* 2025;141:109806.
- [23] Zhang Y, Yu P, Xiao Y, Wang S. Pyramid-structured multi-scale transformer for efficient semi-supervised video object segmentation with adaptive fusion. *Pattern Recognit Lett* 2025;194:48–54.
- [24] Zhang Y, Zhang Y, Xiao Y, Wang T. Spatiotemporal dual-branch feature-guided fusion network for driver attention prediction. *Expert Syst Appl* 2025;292:128564.
- [25] Zhu W, Lu J, Li J, Zhou J. Dsnnet: a flexible detect-to-summarize network for video summarization. *IEEE Trans Image Process* 2020;30:948–62.
- [26] Wang Z, Zhu Y, Wu X. Domain adaptive video summarization using generalized transformer. *Pattern Recogn* 2026;170:112059.
- [27] Chellam JAI, Malliga P, Manoharan M, Ramkumar M. Resnet 152 convolutional adaptive spiral flying sparrow search neural network for image enhancement of video surveillance scene. *Signal, Image and Video Processing* 2025;19(3):214.
- [28] Jeenath Shafana N, Ayothi S. Efficient brain tumor detection and segmentation using DN-MRCNN with enhanced imaging technique. *Microsc Res Tech* 2025.
- [29] Soni LN, Wao AA. Face detection under low-light and low-resolution conditions using contrast-limited adaptive histogram equalization and a modified convolutional neural network. *J Neonatal Surg* 2025;14(32s):2620–31.
- [30] Prabhu MR, Sivaraman R, Nagabhooshanam N, Kumar RS, Salunkhe SS. Empowering artificial intelligence-based multi-biometric image sensor for human identification. *Measurement: Sensors* 2024;33:101082.
- [31] Al Bataineh A, Vamsi B, El-Abd M, Doppala BP. Kolmogorov–Arnold Networks for predicting carotid intima-media thickness in cardiovascular risk assessment. *Sci Rep* 2025;15(1):32108.
- [32] Liu Z, Mao H, Wu CY, Feichtenhofer C, Darrell T, Xie S. A ConvNet for the 2020s. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*; 2022. p. 11976–86.
- [33] Tariq A, Iqbal MM, Iqbal MJ, Ahmad I. Transforming brain tumor detection empowering multi-class classification with vision transformers and efficientnetv2. *IEEE Access* 2025.
- [34] Kaveri K, K VRP, Reddy GP, Pothamsetti SLG, Radha K, Penumaka R. ACO-optimized MobileNetV2-ShuffleNet hybrid model for automated dental caries classification. *Sci Rep* 2025;15(1):40615.
- [35] Aboelenin S, Elbasheer FA, Eltoukhy MM, El-Hady WM, Hosny KM. A hybrid Framework for plant leaf disease detection and classification using convolutional neural networks and vision transformer. *Complex Intell Syst* 2025;11(2):142.
- [36] Duan Y, Song C, Zhang Y, Cheng P, Mei S. STMSF: swin transformer with multi-scale fusion for remote sensing scene classification. *Remote Sens* 2025;17(4):668.
- [37] Shekhawat D, Chaudhary D, Kumar A, Kalwar A, Mishra N, Sharma D. Binarized spiking neural network optimized with momentum search algorithm for fetal arrhythmia detection and classification from ECG signals. *Biomed Signal Process Control* 2024;89:105713.
- [38] Negm A, Ma X, Aggidis G. Deep reinforcement learning challenges and opportunities for urban water systems. *Water Res* 2024;253:121145.
- [39] Lang Y, Gao Y. Dream Optimization Algorithm (DOA): a novel metaheuristic optimization algorithm inspired by human dreams and its applications to real-world engineering problems. *Comput Methods Appl Mech Eng* 2025;436:117718.
- [40] Cao D, Zhou H, Wang Y. Improve the image caption generation on out-domain dataset by external knowledge augmented. *Multimed Syst* 2025;31(1):61.
- [41] Chel YH, Poh LL. Brain Tumor classification in MRI: insights from LIME and Grad-CAM explainable AI techniques. *IEEE Access* 2025.