



STAGE-ORIENTED ANALYSIS OF VIDEO SUMMARIZATION MODELS ACROSS LEARNING PARADIGMS AND BENCHMARK DATASETS

Aruna M^{*1}, Rajiv Ranjan Singh²

¹Assistant professor, Department of Electronics and Communication Engineering, Presidency university, Itgalpur Rajanakunte, Yelahanka, Bengaluru, Karnataka, India.

²Professor and Head of the Department, Department of Electronics and Communication Engineering, Presidency University, Itgalpur Rajanakunte, Yelahanka,,Bengaluru, Karnataka, India.

¹<http://orcid.org/0000-0001-8743-5215>, ²<http://orcid.org/0000-0003-4217-7552>

Email: aruna.m@presidencyuniversity.in, rajivranjansingh@presidencyuniversity.in

ARTICLE INFO

Article History

Received: February 17, 2026

Reviewed: April 26, 2026

Accepted: May 26, 2026

Published: June 30, 2026

Keywords:

Video Summarization,
Stage-wise Analysis,
Multimodal Feature Fusion,
Temporal Modeling,
Relevance Scoring and Ranking.

ABSTRACT

Video summarization is to create concise summaries for longer videos. Although many researchers contribute with great advancements in summarization process, there are also significant obstacles to overcome. One prime challenge is the subjective nature of determining the important elements in the video based on the viewer. Another major issues are due to the various stages in a traditional pipeline, where an error at one level can result in failure of the entire summarization process. In addition, many current methods unable to retain connections between elements throughout the videos, and scalability continues to be an obstacle for the creation of lengthy summaries. This review categorizes video summarization process into stages, which include video preparation and segmentation, extracting features from different sources like visual, audio or text features, modeling patterns based on time, scoring for relevance, and finalizing the summary. Traditional approaches relied on basic machine learning models, where the modern approaches use advanced deep learning and attention-based methods that can better capture global context and semantics of a video. Challenges still exist, such as high computational costs, heavy reliance on human annotation and having a bias in the use of data sets. This survey compiles evidence from recent benchmark studies to better understand transformer based self-supervised and reinforcement learning techniques can address the limitations of traditional approaches but also introduce new concerns related to computation complexity and interpretability. Furthermore, this review provides future research directions to achieve an integrated, end-to-end, and user-preferred system and conclude that multimodal reasoning as a key enabler for ensuring high-quality, scalable, and semantically consistent systems for video summarization applications



Copyright ©2026 by authors and Galileo Institute of Technology and Education of the Amazon (ITEGAM). This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

I. INTRODUCTION

With an increasing amount of video files through surveillance, social media, streaming services and smart homes, managing that content has become a larger issue than ever before. As a result of its huge size, length and complexity of the video content, traditional methods of browsing and retrieving video are no longer feasible. Though human inspection is valuable it is very slow, highly subjective, and mentally demanding. Which results in many inconsistent outputs when inspecting for large systems [1]. So, it is essential to utilize video summarization to take long videos and create shorter, more informative summaries of those videos, that still capture the most important information from those long videos. With its strong potentials, the AI-based summarization framework helps to improve retrieval efficiency, decrease storage requirements, and assist with tasks such as recommendations, monitoring, and multimedia analytics [2], [3]. Significant advancements such as offline key-frame extraction to context-conscious and task-specific summaries are better able

to reflect both the semantic meaning of an event and the needs of users [4], [5]. These advancements place a high value on systematic reviews to evaluate both the performance of algorithms and the design elements that influence the video summarization process.

Over the years, video summarization has undergone a significant transformation from simple rule-based approaches to advanced deep learning and multimodal solutions. The earliest methods used shot boundary detection, clustering and hand-crafted visual features to identify representative frames, as these methods were able to generate summaries very quickly. However, they struggled to represent video with diverse content and could not capture temporal relationships across a lengthy period. Traditional machine learning models like Support Vector Machines (SVM), Hidden Markov Models (HMM), Sparse Coding provide better consistency through frame importance; however, they also strongly relied on hand-designed features and could not capture long-range dependencies [6], [7]. Deep learning has transformed the landscape. CNN-RNN pipelines, attention mechanisms, and GANs have enabled end-to-end learning from raw video directly to importance scores [8] [9]. Transformers and multimodal networks take this a step further by being able to model long-range, temporal dependencies and coordinate visual, auditory and other contextual signals [10-12]. In addition to the above techniques, recently developed vision-language models rely on external knowledge and cross-modal reasoning to produce summaries that better match human understanding of perceptions and intentions [13], [14]. Despite these advances, many studies examine video summarization systems as a single system, limiting the understanding of the individual video summarization stages' contributions to performance. A more focused, stage-by-stage analytical approach is needed for detailed analytical reviews.

Video summarization not only needs overall evaluations but also needs to be analyzed as individual stages to show how design decisions made at each stage affect both performance and generalizability. Each stage consists of pre-processing and segmentation, feature extraction from multiple types of media, temporal modeling of relevant video sequences, and post-processing, which is further divided into selecting summaries, smoothing, and so on [15]. Evaluating individual stages provides the ability to identify where improvements occur and where the bottlenecks in video summarization systems exist, whether from heuristics, machine learning, deep learning, or hybrid approaches. This approach is crucial when comparing models across many different benchmarks, namely SumMe, TVSum, VSUMM, YouTube Highlights and ActivityNet which all differ in length, complexity, annotation style and subjectivity [16]. Challenges persist despite progress; most modern-day video summarization systems still encounter practical limitations in the robustness and usefulness of their performance in the real world. Modular pipeline architectures also tend to suffer from cumulative error propagation; which lead to incorrect relevance scores in later pipelines. Furthermore, subjectivity in annotation and bias in datasets limits the ability for fair evaluation in common benchmarks and the generalization of results to other video types. In addition, current models tend to suffer from balancing long-range temporal dependency modeling with computing efficiency for both long-form and multimodal videos. Overall, these types of problems indicate an absence of a unified analysis framework that is capable of systematically illustrated video summarization methodologies at all stages of processing while considering issues such as scalability, semantic consistency, and interpretability. This analysis makes several important contributions as follows,

1. A structured end-to-end framework for video summarization is proposed that divides the relevant stages into five major processes as Preprocessing and segmenting, extracting multimodal features, modeling temporal relationships, scoring relevance, and postprocessing. All these processes in this framework are analyzed and evaluated through comparative methodology.
2. Three different methods such as machine learning, deep learning and hybrid approach, are critically reviewed using a series of benchmark datasets to identify and illustrate differences in performance, scalability, and reliability.
3. This review considers all of the open issues such as annotation bias, effect of errors propagating through systems, the difficulty of modelling long sequences, with all computational restrictions. A common viewpoint has therefore been taken on each of these areas and discussed on their relationship with different paradigms.
4. Future research directions towards unified, multimodal, and preference-aligned summarization systems, in particular transformers, self-supervised learning algorithms and reinforcement learning methods are thus identified as a prominent way in developing the frameworks for future systems.

Research themes over the period from 2020 to 2025 are illustrated in Figure 1, where (a) shows the trends in the datasets adopted since 2020, while (b) depicts the evolution of models from classical machine learning (ML) techniques to vision-language models (VLM).

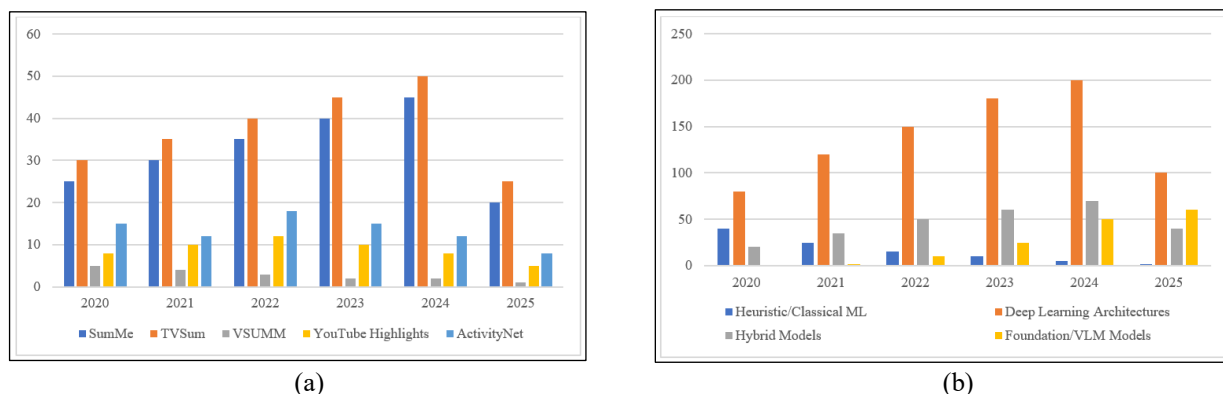


Figure 1: Benchmark Dataset Usage and Architectural Trends Over Five Years.

Source: Authors, (2026).

This review has been conducted in accordance with methodological rigor and transparency as outlined by the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines for literature identification, screening, eligibility assessment and inclusion.

A worldwide systematic search has been conducted from the principal scientific databases including IEEE Xplore, Scopus, Web of Science and ScienceDirect using common structured keyword combinations related to video summarization, multimodal learning, temporal modeling, and relevance scoring. After removing duplicates, the title and abstracts were screened to identify non-peer-reviewed journals, short abstracts and studies not related to video summarization. The full-text version of any remaining source was assessed for the methodological contributions made to the research and the empirical evaluation of the benchmark dataset or comparison across different stages of the summarization process. Studies published between 2020 and 2024 are preferred so that the most current advancements were represented while allowing for the retention of seminal works that provide context for the current research advancements. The final set of studies selected for this review includes a relative balance of the different research paradigms including heuristic methods, machine learning, deep learning, hybrid methods and foundational model-based methods. This PRISMA-aligned selection process, in combination with the balanced representation of heuristic and machine learning approaches, is designed to minimize potential biases due to selection processes, increase reproducibility, and thereby enhance the validity of the comparative insights presented within this review. Figure 2 presents the taxonomy that has been established from this review as a result of the PRISMA-guided processes.

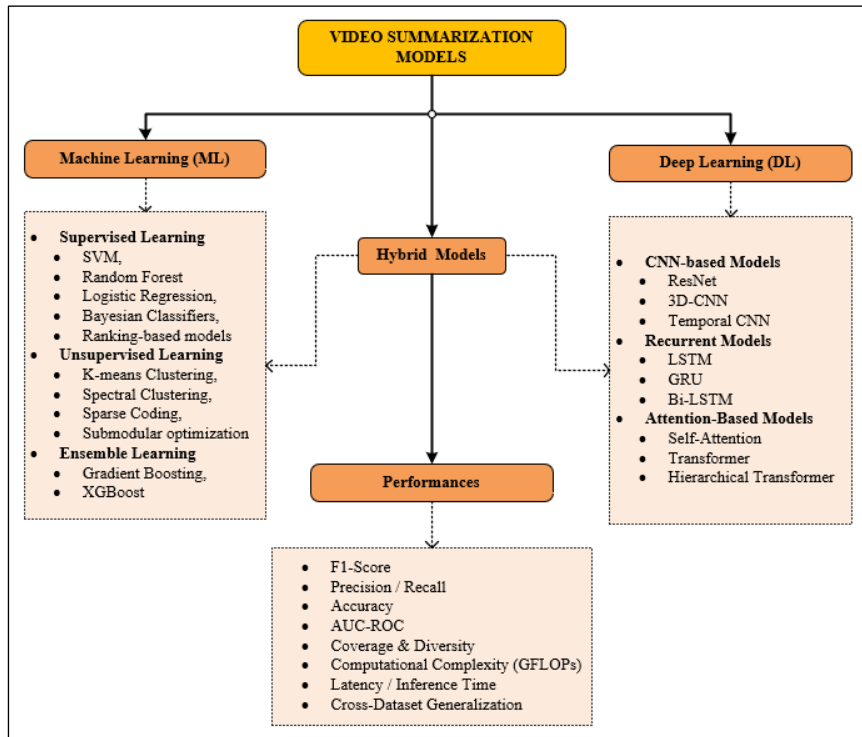


Figure 2: Taxonomy of the Literature Review.

Source: Authors, (2026).

The organization of the analysis is such that: Section 2 describes the methodology based on its stages; Section 3 provides a comprehensive analysis based on different models and its corresponding data; Section 4 describes the key insights, challenges, and implications of the study; Section 5 outlines the future state and research opportunities; and Section 6 provides conclusion remarks and future research directions.

II. METHODOLOGY

The video summarization model shown in figure 3 is a large-scale video summarization system that has well-defined stages through which the functionality of the video is extracted and reduced in size while keeping its semantic content. The first stage is the pre-processing and temporary segmentation of the raw video into shots or events with cleaning, noise reduction, and boundary detection of the video which results in a reduction of redundancy and improved structure. The second stage is extracting multimodal features where visuals, motion, audio or text is processed using various techniques such as handcrafted descriptors or deep neural networks. These multimodal features are fed to a temporal modeling module in order to develop sequential patterns and learn long range relationships through a variety of models such as HMMs, Long Short-Term Memory (LSTM) networks, Gated Recurrent Units (GRUs), transformer algorithms, hybrid algorithms. The third stage is to apply a relevance scoring mechanism to each frame or segments of the original video based on rules, supervised or unsupervised learning methods, attention mechanisms, or reinforcement learning. The last stage is using various post-process methods like top-K selection, thresholding, redundancies removal, and smoothing to result in a concise and coherent summary of the original video while maintaining continuity and diversity.

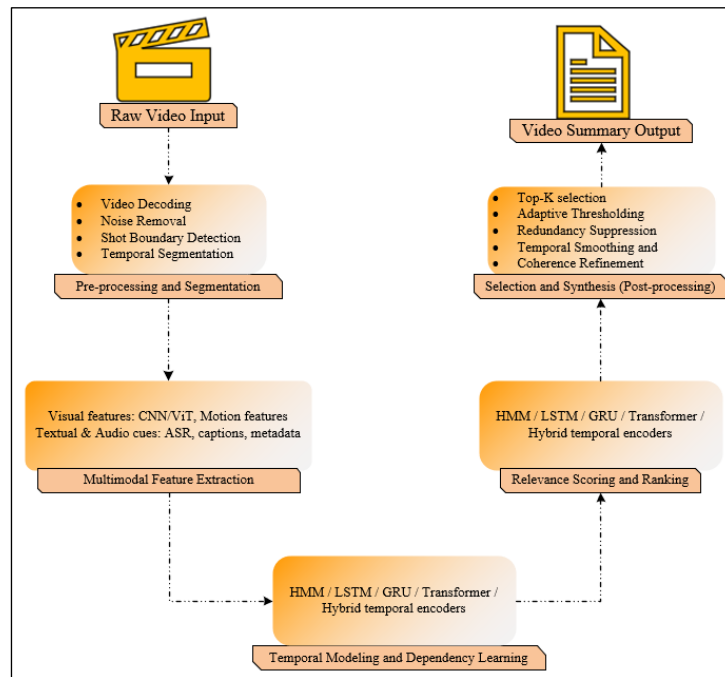


Figure 3: Block Diagram for Different Stages in Video Summarization Process.
Source: Authors, (2026).

An analytical framework based on the stages is described in this section to analyze video summarization algorithms evaluated using different methods like segmentation, feature extraction, temporal modeling, relevance scoring and post-processing. These methods include heuristic, ML, DL, and Hybrid types of algorithms. They have all been compared against each other through the use of common benchmark datasets in order to evaluate their performance, reliability, and scalability.

II.1 PRE-PROCESSING AND SEGMENTATION

Video summarization involves a lot of pre-processing and segmentation of videos before they are summarized. This includes efficient summary generation, there redundancy, its coherent; therefore, pre-processing and segmentation will create a summary that feels coherent, does not contain redundancy, and runs efficiently. Of primary importance in this process is shot boundary detection, which separates a video into segments that have meaning. Using modern deep learning technologies such as convolutional neural networks (CNNs), researchers have been able to achieve nearly perfect accuracy (approximately 98%) in detecting both abrupt and gradual transitions between video frames (i.e., the transitions from one frame to another), thus allowing researchers to reliably detect the most important shots [17]. Additionally, using shot detection in conjunction with clustering and optimization methods will increase the quality of keyframes extracted from a video and improve the performance of key management processes using the benchmark datasets [18]. In specialized situations, video lectures and hybrid audio-visual segmentation techniques that utilize optimization algorithms and color models have been shown to achieve more than 90% accuracy according to structured settings [19]. The other important aspect is the frame sampling, adaptive sampling methods that adapt their sampling density by either motion or semantic richness have been shown to outperform standard sampling methods and retain only those frames that contain the most information, discarding those that contain less information [13], [20]. In addition to adaptive sampling, using dimensionality reduction and clustering techniques to eliminate redundant data and reduce processing costs without losing the integrity of a video summary provides an additional benefit [21-24]. Together, these methods enable scalable and accurate summarization, even on challenging datasets such as SumMe and TVSum, where diverse content and subjective annotations make evaluation difficult [25].

II.2 MULTIMODAL FEATURE EXTRACTION

Multimodal feature extraction is regarded as one of the key advancements in video summarization process practiced today, as it allows for an improved semantic space representation and an increase in the relevance of summaries produced. Researchers have used a variety of available combinations (e.g., visual-only; audio/visual; vision/language) to determine which combination would produce the best representation of visual content that conveys meaning. Numerous studies have indicated that by combining visual descriptors with audio descriptors (e.g., color histograms, motion-related descriptors and object category-based descriptors with audio descriptors such as spectral flux and energy pattern) produce stronger results than when each type of descriptor is used independently. Further, the use of multimodal descriptor fusion strategies provides for improvements in terms of evaluation metrics such as AUC and F1-score, supporting the complementary nature of audio and visual information used in video summarization tasks [6]. As demonstrated in Figure 4, new frameworks incorporating lexical attributes as well as outside knowledge [25]. Vision-language fusing allows for reasoning at higher semantic levels and allows for estimating the importance of a piece to a topic more accurately. Fusion models have the forerunners of the visual model that are pure visual-based and therefore, they outperform purely visual methods in both synthetic or real-world settings [26]. The literature shows that handcrafted features (often utilized with traditional classifiers) are both efficient and interpretable; however, they are restricted by their inability to model complex spatio-temporal semantics (i.e., relationships) [27].

Conversely, feature representations derived from machine learning methods such as CNNs, RNNs, and hybrid deep networks provide richer spatial and temporal feature representation leading to increased robustness and greater accuracy across datasets [28]. Because of this, it is common to see the fusing of both handcrafted/learned feature representations to reduce redundancy and increase segmentation quality ultimately improving classifier performance through ablation studies [29]. Therefore, multicasting/learning-based features are typically richer semantically and materially in terms of providing context and require larger computational resources.

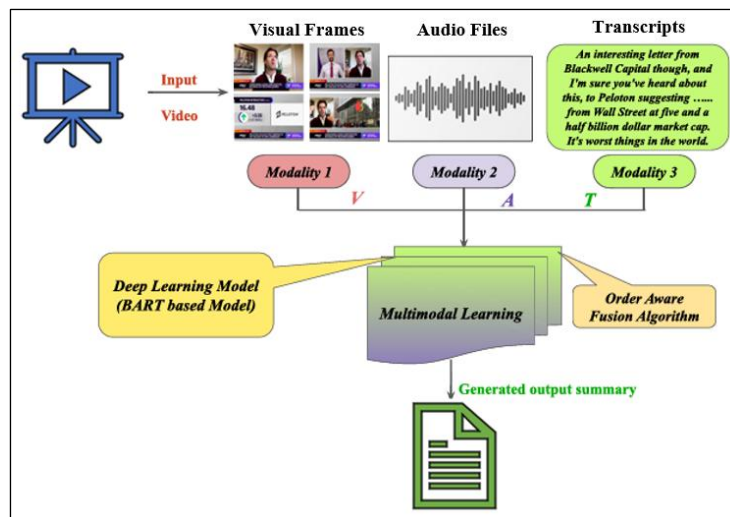


Figure 4: Multimodal Summarization Workflow.

Source: [26].

II.3 TEMPORAL MODELING AND DEPENDENCY LEARNING

Temporal modeling forms the basis of producing clear and high-quality summaries of videos, but different types of temporal models will produce summaries of videos with different characteristics. HMMs is often used when the ability to interpret the summary and the need for a small amount of data are important. It is a good way to represent the short-term transitions that take place in the video; however, since an HMM represents short time (i.e., it has a Markovian character) it cannot represent the complex long-term dependencies that might exist in the video. Deep recurrent networks, such as LSTM networks and GRU, can be implemented to overcome the limitations of modeling long-term dependencies with HMMs by implementing a gating mechanism that maintains temporal context over long sequences of time. While LSTMs provide an excellent representation of long-term dependencies in time series data, they are computationally inefficient. In contrast, GRUs produce similar-quality presentations of long-term dependencies, and they do so more efficiently than LSTMs do, particularly when it comes to multimodal and large datasets [30]. The use of self-attention in transformer-based architectures increases the ability to model temporal relationships globally without recurrence and allows for the simultaneous processing of these relationships. Thus, transformer-based architectures eliminate the limitation imposed by recurrent networks, which must model sequences in a sequential manner [31]. The computations needed to run a standard transformer on a long sequence are typically very high; however, there are innovative designs, such as hierarchical transformers and memory-augmented transformer architectures, that provide an efficient means of learning from sequences with long-range dependencies. Attention-based architectures consistently out-perform HMMs and stand-alone recurrent architectures in empirical evaluations, so they have become the dominant architecture in modern video summarization [32], [33]. Figure 5 shows the core DL model that uses visual and language models for video summarization.

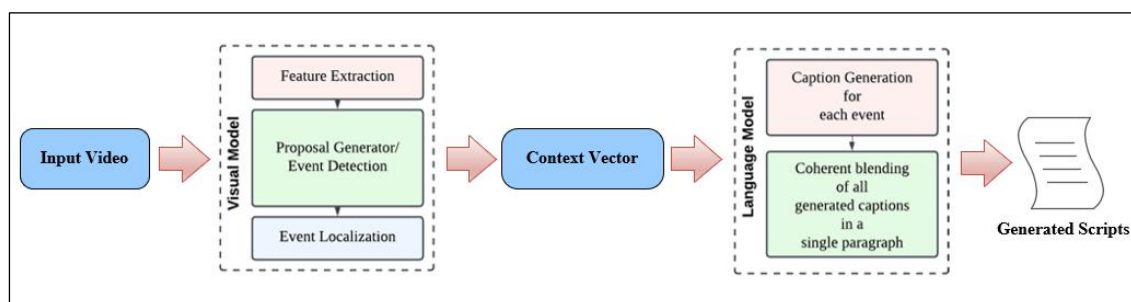


Figure 5: Fundamental Video Description Model.

Source: [33].

II.4 RELEVANCE SCORING AND RANKING

Video summarization relies heavily on scoring and ranking to choose which frames or segments will appear in the summary. Traditionally, scoring and ranking were based mostly on the relevance of a frame, using very simple but effective heuristics to define how important the frame when compared to other frames. Commonly used sources for heuristic scoring are appearance, motion, and semantic cues; local dissimilarity and global consistency are often pairwise metrics.

Heuristic approaches to frame retrieval typically use a combination of deep feature extraction and reward-based reranking to strike a balance between relevance and diversity, with satisfactory performance even for training-free or zero-shot scenarios. [34]. In figure 6, the use of frame, semantic space and color coding of the video content are conceptually demonstrated, showing how local dissimilarity, global consistency and uniqueness might be identified in a video using video features. Supervised learning approaches can predict frame/segment importance scores based on human-annotated data and achieve high F1-scores while also semantically aligning with what users expect. However, their dependence on labeled data limits their ability to scale. In contrast, there are some subsequent methods that exploit both reinforcement learning and adversarial methods to find the best possible reward function that honors time continuity, representativeness and diversity within the response, and also has shown performance that is comparable to supervised performance across many benchmark datasets [15], [35]. Additionally, these same approaches can be enhanced with the use of attention mechanisms that leverage attention to account for both local and global temporal context. Finally, reinforcement learning provides reward signals that provide guidance for ranking towards uniformity and semantic richness without the need for explicit supervision [36], thus hybrid scoring approaches which utilize heuristic cues, attention-based learning and reinforcement reward criteria consistently outperform either purely heuristic or purely supervised approaches, thereby providing a scalable, robust approach to modern video summaries.

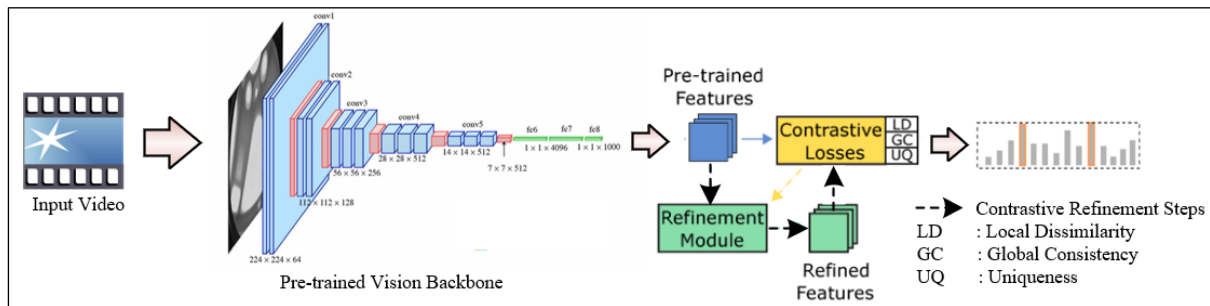


Figure 6: Leveraging Contrastive Learning for Video Feature Analysis.

Source: [34].

II.5 SELECTION AND POST-PROCESSING

Selecting and editing clips, the final part of creating a summary video, results in creating a final product that is coherent, enjoyable, and dynamic. Using smooth transitions between different scenes makes the presentation appear seamless throughout. Advanced technologies are being developed to allow for greater flexibility in the initial phases of editing such as temporal convolution networks [37], coarse to fine scale models [38], and use of graph-based temporal aggregation methods [39]. Temporal security will remain intact by using memory optimization through temporal security, motion anomaly indices, dimensional percentage decrease methods, and intelligent cluster processing—resulting in a video summary that is both short but representative of the original raw video footage. In addition to providing more diverse representations of video material, these approaches will also lower the computational costs associated with creating summaries, which is an important consideration in long or specialized videos where repeated patterns occur very often [23], [39], [40]. Adaptive thresholding and Top-K ranking are mechanisms for selecting components for a summary. Adaptive methods are the best choice because they allow the number of selected frames or segments to be adjusted based on learned importance scores, length and complexity of the video. By adapting to current conditions, and by avoiding over-summarization (which has the potential to miss or lose critical information), and under-summarization (which may create redundancy), adaptive methods help reduce both redundancy and loss of detail in the summary. Research indicates that adaptive selection outperforms fixed thresholds in achieving balanced coverage to conciseness ratios for different datasets and genres [41], [42]. Temporal smoothing, redundancy suppression, and adaptive selection are three complementary processes that enhance coherence, informativeness, and generalization when creating a video summary.

III. ANALYSIS

This section provides an overview of various approaches to summarizing videos using heuristics (i.e., rules of thumb), ML, DL, or hybrid methods. It evaluates video summaries using performance trends, bias in datasets, computational complexity, generalization ability, and scalability by comparing established benchmark data sets and drawing conclusions about their strengths and weaknesses.

III.1 STAGE-WISE COMPARATIVE ANALYSIS

Examining a stepwise approach to video summarization using ML, DL and Hybrid approaches reveals that, there are clear strengths and weaknesses to each approach at each stage in the process. At the segmentation stage, ML techniques mainly use handcrafted video descriptors and classical classification approaches, making it fast, interpretable and relatively easy to implement. But one of the major limiting factors is adaptability to a diverse or complex video content because of the pre-defined features. In contrast, DL techniques use Convolutional Neural Networks and Object Detection algorithms to automatically generate lists of scenes or events from the source of the video which yields higher accuracy and higher robustness, but require significantly more computational power and longer training times [21], [43]. ML methods focus on identifying engineered audio and visual features that are lightweight, easy-to-understand, but not very deep in their representation. In contrast, DL architectures allowing for the learning of hierarchical spatial-temporal feature representations directly from the data (e.g., CNNs, ResNets, and transformer-based models) demonstrate much richer semantics and context than those found using ML methods [21]. There are also differences due to separate temporal modeling methods; ML methods typically use statistical models or simple sequence-based analyses, but DL methods leverage recurrent architectures (e.g., LSTM or GRU) and attention-based transformers to identify long-range dependencies.

Hybrid frameworks leverage both of these strengths owing to their robustness in scenarios where little training data is available or where resources are limited; thus, they are of great value in practical applications [44], [45]. ML techniques typically employ supervised classifiers or ranking functions to assign relevance scores while DL models use end-to-end methods or reinforcement learning for determining the importance of various elements of videos. This has allowed DL models to adapt much more rapidly than ML models to the variety of different types of video. Hybrid methods use an ML-based approach for selecting candidate video segments for inclusion in a summary, followed by the use of DL methods to optimize those selections, which results in adaptive, context-sensitive importance estimates that can balance efficiency and accuracy [6], [44]. Post-processing also demonstrates these two groups of algorithms have distinctive characteristics, a ML provides clustering and filtering functions to increase coherence, while DL contains adaptive thresholds and temporal interpolations that provide for smoother transitions, and hybrid methods refine DL results through ML redundancy suppression [46]. Overall, DL methods provide superior representation learning and temporal modeling abilities, while ML provides for greater efficiency and interpretive capabilities than DL; however, Hybrid approaches have resulted in the most efficient and generalizable performance across benchmark datasets making Hybrid approaches the most effective choices in real world video summarization scenarios. Table 1 summarizes a comparative discussion of all these methods.

Table 1: Comparative Analysis of Different Video Summarization Approaches.

Pipeline Stage	ML Methods	DL Methods	Hybrid Methods
Segmentation	Handcrafted features & classical classifiers; efficient but less adaptable [44]	CNNs & object detectors; automated & accurate but computationally expensive [21]	Combines ML efficiency with DL accuracy [44]
Feature Extraction	Engineered visual & audio descriptors [43]	CNNs, ResNets, & transformers; learns complex semantics [21]	Integrates ML descriptors with learned representations [44]
Temporal Modeling	Simple statistical or sequence models [44]	LSTM, GRU, & transformers; captures long-range dependencies [6]	Combines strengths to enhance robustness [44]
Relevance Scoring	Supervised classifiers [6]	End-to-end predictions or reinforcement learning [6]	Integrates ML-driven selection with DL optimization [45]
Post-processing	Clustering & filtering for coherence [44]	Adaptive thresholding & interpolation [44]	Refines DL outputs with ML-based redundancy control [46]

Source: Authors, (2026).

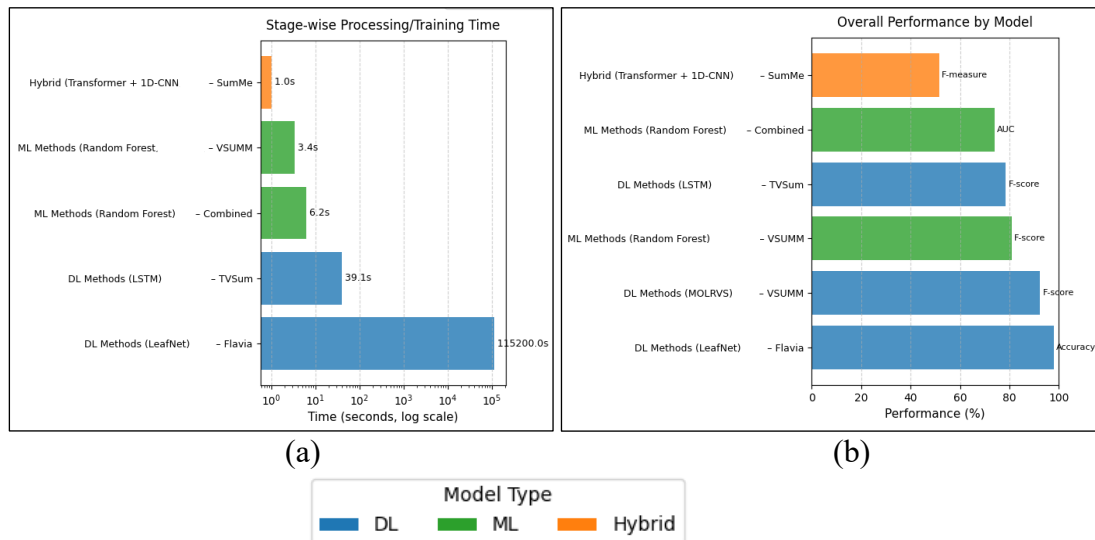


Figure 7: Stage-wise Benchmarking of Video Summarization Models.

Source: Authors, (2026).

The comparative analysis in Figure 7 shows a clear distinction between the cost of computing versus effectiveness in video summarization/classification models using a stage-based approach. Figure 7(a) shows a significant time difference between LeafNet's training time of 32 hours in approx. (but only ~7 seconds for hybrid/ML) indicating cost, efficiency and research trade-offs as shown in Figure 7(a). Similarly, Figure 7(b) shows varying performance based on task and metrics; although DL has high F-scores or accuracy (up to 97.9%) compared to other models, the latency of ML is much lower but still provides similar-quality outcomes. Additionally, the dual-panel chart compares both time on a log scale across performance without conflating different benchmark types when making deployment decisions especially for real-time or limited resources.

III.2 DATASET-CENTRIC PERFORMANCE ANALYSIS

The evaluation of video summarization through dataset centric evaluation demonstrates how the size or volume of datasets (e.g., their design, their annotation protocols and the inherent biases associated with those datasets) impact how model performance is rated. One common form of short-style dataset, like SumMe and TVSum, contains human-annotated videos of usually fewer than a couple minutes long are used as standard examples to evaluate a model's ability to select keyframes, determine accounting relevance and provide visual diversity. Such datasets allow for controlled experimentation and consistent benchmarking across each other, but because they provide limited temporal scope to the models, the models rarely encounter complex long-range dependencies which are required for generalized summarization [5]. The subjectivity behind human annotation of importance scores (i.e., the bias associated with how a human may rate a video as being important vs. another rating it as less important), will also lead to overfitting of annotation styles and therefore models that derive high F-scores & precision within these datasets at times will be unable to perform adequately on unseen real-world content [18], [21].

In contrast, while YouTube Highlights focuses on event centric content, where the focus is on parts of the event that are of higher activity, and will therefore be more beneficial to the detection of highlights, this focus will cause a potential over-estimation of the capability of the model due to the lower representation of those types of events and activities and will therefore decrease the ecological validity [43]. ActivityNet, however, consists of longer unchanged videos and overlaps between the different events, as well as a complex temporal structure. Models must be able to model long-term dependencies, semantic progression and continuity of context [47]. Demonstrating success on these types of datasets is an indicator of real-world applicability, scalability and robustness. As a result, the findings show that strong performances on short-form or event-based datasets will not predict success on long-form or diverse video datasets. Therefore, dataset-aware evaluation is essential to ensure benchmark selection aligns with research goals, reflects diverse temporal and semantic challenges, and accurately represents practical summarization capabilities [43], [47], Table 2 summarizes the characteristics, biases, and performance implications of different datasets.

Table 2: Comparative Analysis of Dataset-Based Performance Analysis Approaches.

Dataset	Characteristics	Biases	Performance Implications
SumMe	Short-form videos; human-annotated summaries	Human-annotated bias; subjective preferences	Strong F-score and precision in short videos, may struggle with generalizability [5], [21], [48]
TVSum	Short-form videos; human-annotated summaries	Similar to SumMe	Emphasizes semantic relevance and visual diversity; limited temporal dynamics [5], [21]
YouTube Highlights	Event-centric; focuses on short, clearly defined highlights	Event-centric bias; favors discrete activities	May overestimate model performance; emphasizes activity detection over subtle content [44]
ActivityNet	Event-centric; focuses on short, clearly defined highlights	Complexity of temporal dependencies	Tests models on capturing long-range dependencies and nuanced transitions; emphasizes scalability [47]

Source: Authors, (2026).

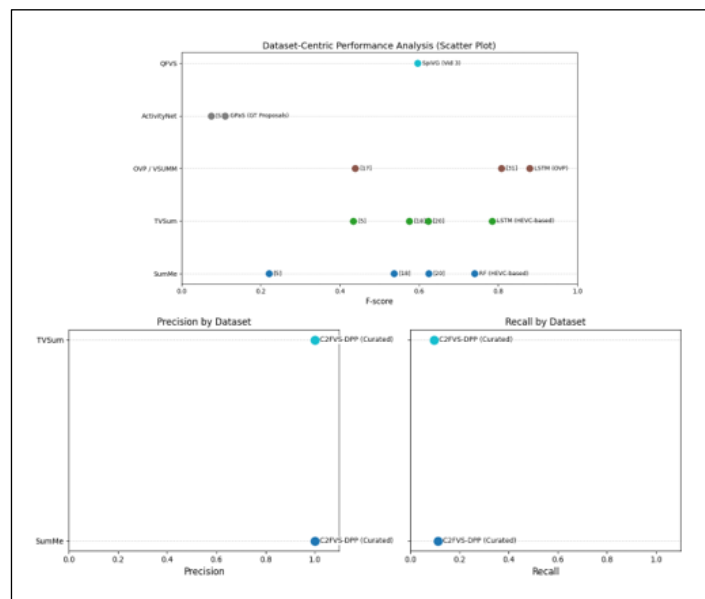


Figure 8: Dataset-Centric Comparison for Video Summarization Methods.

Source: Authors, (2026).

In Figure 8, a scatter plot comparing Video Summarization Models across five datasets is shown along with the models that performed best on each of the datasets. Each point corresponds to a different model, such as C2FVS-DPP (Curated) [5], Highlight Detection [49], SpiVG [48], MMAN [50], RF (HEVC-based) [21], LSTM (HEVC-based) [21], GPaS (Automatic Proposals) [47] only the best is labeled to avoid cluttering.

The plot shows that there is vast variability in performance within the different benchmarks where the Youtube Highlights Dataset performed the best and the ActivityNet Dataset performed the worst showing the inherent differences in difficulty among each of the datasets.

III.3 COMPUTATIONAL COMPLEXITY AND SCALABILITY

Video summarization entails a challenge of computational complexity and scalability; therefore, an assessment of the computational efficiency of lightweight ML models relative to heavyweight transformer-based architectures is fundamental. For example, lightweight ML models excel in their ability to execute with computational efficiency, as they require fewer parameters, smaller numbers of floating point operations (FLOPs), and use either simplified attention or convolution to deliver results quickly (ie; in real-time or at the edge) in environments where available resources are limited [51]. Conversely, transformer and vision-language models offer superior representational ability through the representation of complex temporal and contextual dependencies at the cost of extensive computational requirements (ie; high FLOPs, extreme memory requirements and delays), which create scalability issues for those platforms with limited resources [44]. Video summarization involves balancing high-quality summaries from large Transformer models and fast, economical inference of smaller-sized ML models for short or moderately complex videos [52]. In response, various research efficiencies have emerged, such as token sparsification, probabilistic pruning, hierarchical tokenization, which reduce computing costs while producing the bulk of the performance [53]. Hybrid designs, incorporating convolutional layers and attention modules have reduced, but still use computationally expensive, self-attention mechanisms. In addition to using knowledge distillation and quantization to improve the efficiency of deployment without significantly impacting accuracy [54], lightweight transformer variants have provided an additional means of producing real-time summaries for all types of activities while yielding very complex, long-form videos; allowing for rapid delivery of scalable, near real-time summaries across many different types of activity [55]. Table 3 shows the trade-offs in accuracy and real-time feasibility.

Table 3: Trade-offs in Accuracy and Real-Time Feasibility of Video Summarization Techniques.

Aspect	Lightweight ML Models	Heavy Transformers
Computational Complexity	Significantly lower due to fewer parameters and simplified operations [51]	Higher due to extensive floating-point operations (FLOPs) and large memory footprints [52]
Scalability	Highly scalable for resource-constrained devices; suitable for real-time applications	Limited scalability; struggles on resource-constrained devices [44]
Example Models	Lightweight transformers (e.g., TBiGNet), CNNs with simplified architectures	Standard Vision Transformers, large language models
Accuracy	Competitive accuracy; over 99% on short videos; may sacrifice some representational power	Superior representational power; high performance on complex tasks
Real-time Feasibility	Faster inference and lower resource consumption, enabling real-time summarization	Higher latency hinders real-time deployment
Optimization Strategies	Employ sparsity, depthwise separable convolutions, efficient attention mechanisms [53]	Token sparsification, Bayesian prototypical pruning, Neural ODEs for temporal localization [54]
Trade-Offs	Good accuracy with real-time feasibility; may limit performance on complex videos	High accuracy but at the cost of computational overhead and latency [55]

Source: Authors, (2026).

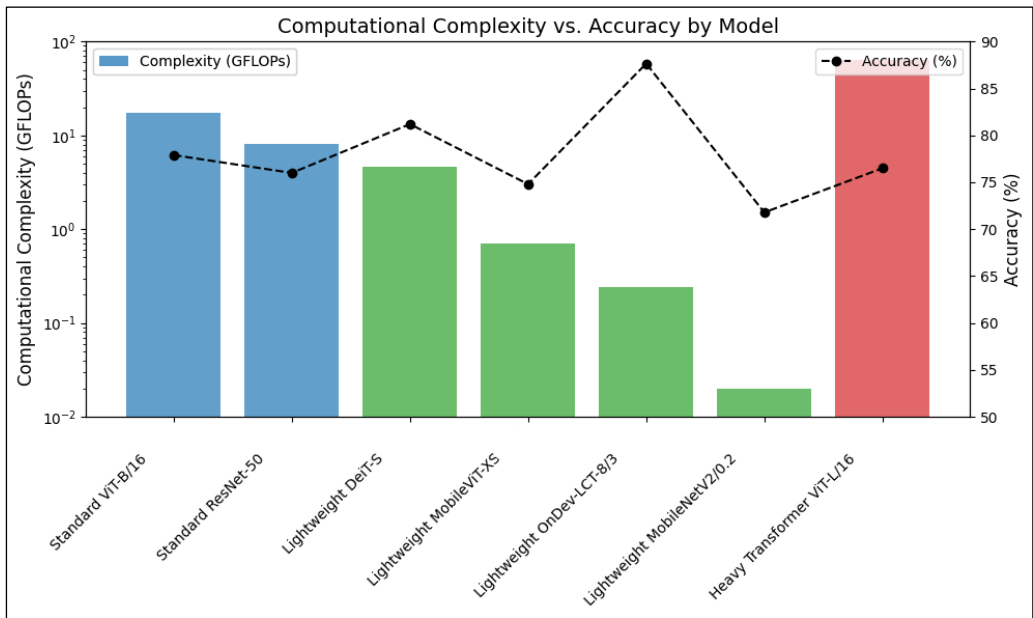


Figure 9: Computational Complexity Analysis Across Recent Models.

Source: Authors, (2026).

Data in Figure 9 shows how computational complexity of the vision models (in GFLOPS) can be correlated against accuracy and demonstrate the significant trade-offs between model efficiency and model performance. The computational complexity is presented on a logarithmic scale and therefore allows a comparison of both the range of ultra-lightweight MobileNetV2 (0.02 GFLOPS) through to ultra-heavy ViT-L/16 (63.6 GFLOPS) on one graph while accuracy is displayed using a linear scale on the secondary-axis. Ultra-lightweight models (OnDev-LCT-8/3), provide a combination of low resource requirements with high accuracy (87.65% accuracy) as compared to many larger models that vary in architecture (e.g., ViT-B/16 and ResNet-50). Whereas many traditional architectures provide balanced performance metrics; ultra-heavy transformer architecture reveal diminishing returns regarding operationally constrained model selection by resource limitations.

III.4 GENERALIZATION AND ROBUSTNESS

Video summarization models must possess both generalization and robustness because they are crucial characteristics that dictate the degree of an individual model's ability to perform across different datasets and types of videos. A video summarization model will exhibit domain shift resilience due to variations in the content style, how it is annotated, and the semantics being applied, when that model is stable across a number of distinct benchmark datasets [18]. Models that are built using large-scale, heterogeneous datasets like VideoXum, will promote cross-dataset generalization in that they will be able to learn more robust multimodal features and more generalizable semantic relationships [56]. Recent advances in model designs, such as vision-language encoder/decoder frameworks, graph-based reasoning, and variational inference, have improved robustness by capturing more complex temporal dependencies and reducing overfitting to one dataset [47]. Models that do not have adaptive mechanisms also suffer from loss of temporal consistency when used with shorter or longer videos [2], [44]. To address this issue, researchers have proposed new techniques to help with this issue including importance-score interpolation, hierarchical temporal sampling, and reinforcement learning-based methods that reward algorithms for achieving/maintaining temporal coherence [32], [53]. Furthermore, using varied types of data while training by applying semantic conditioning and multimodal fusion, strengthens model robustness in the event of encountering various types of video content and real-world variation [6], [57]. As a result, these approaches lead to scalable and robust video summary systems that consistently perform across many different dataset types (e.g., length) and content platforms [5], [58]. The cross-dataset performance and sensitivity analysis are shown in Table 4.

Table 4: Cross-Dataset Performance and Sensitivity Analysis in Video Summarization Models.

Aspect	Description	Strategies for Improvement
Cross-Dataset Performance Stability	Reflects a model's ability to generalize across diverse video domains and datasets	Leverage large-scale, diverse datasets like VideoXum for richer semantic learning [56] Use advanced architectures (e.g., vision-language encoder-decoder frameworks, graph-based reasoning) to capture complex dependencies [47] Employ variational inference to manage uncertainty in multi-channel features. Test and fine-tune on multiple benchmark datasets (e.g., TVSum, SumMe) [57] Apply semantic conditioning to focus on relevant content
Sensitivity to Video Length	Affects the model's ability to maintain temporal consistency across varying durations	Utilize interpolation methods to adjust predicted importance scores. Combine transformer and CNN architectures for capturing global and local contexts Implement temporal consistency reward functions in reinforcement learning [43] Use hierarchical sampling strategies for fine-grained temporal feature extraction [22]
Content Diversity	Enhances generalization by exposing models to various video types and semantic contexts	Incorporate diverse datasets (e.g., VideoXum) to learn richer correlations [53] Use semantic conditioning for more representative summaries [57] Employ multimodal feature fusion and ensemble learning techniques [6] Implement lightweight frameworks and personalization strategies for diverse outputs [5], [58]

Source: Authors, (2026).

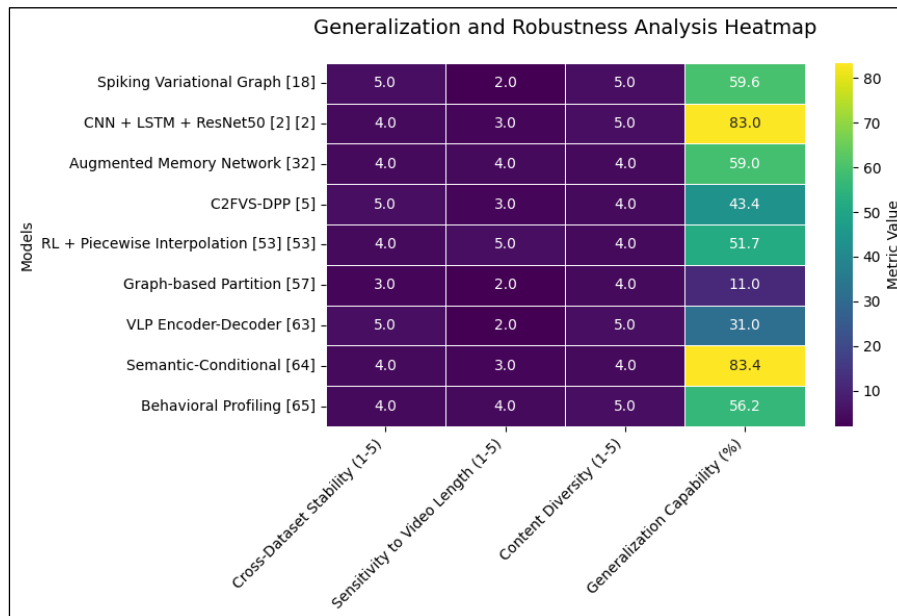


Figure 10: Evaluating Robustness and Generalization across Various Video Summarization Models.
Source: Authors, (2026).

Figure 10 shows the heatmap which illustrates the performance of various video summarization architectures based on stability, sensitivity to video length, and diversity of content using a scale of 1 (low/fragile/niche) to 5 (high/robust/diverse). Notably, both SC-Net and CNN+LSTM achieved more than 83% generalization while maintaining balanced robustness across all stability measures. Both SpiVG and BLIP were uniquely positioned with excellent cross-dataset stability and diversity of content due to their use of hierarchical and vision-language representations. Their low sensitivity to video length also indicates a significant degree of temporal modeling capability. In contrast, the training-free methods C2FVS-DPP leverage features learned during other tasks (i.e., pre-trained features) for exceptional stability without the risk of domain overfitting. In contrast, reinforcement learning approaches are generally more sensitive to video length, presenting difficulties for establishing consistent performance over longer periods of time. The heatmap's color intensity encoding demonstrates these trade-offs clearly; showing how certain choices have impacted adaptability and reliability of their structure in different video environments across a wide range of videos.

IV. DISCUSSION

The comparative analysis of this review indicates that, video summarization research has gone through a significant transformation with regards to the technique from modular pipelines that were initially converged-based to a purely semantic and adaptive unified architecture in the summarization pipeline. In most phases of this pipeline, DL has outperformed both traditional ML and heuristic-based ML in terms of semantic relevance, temporal coherence and robustness. At the same time, there is a significant amount of variability in the performance of the different methodologies when compared against each other based on data set, operational constraints and characteristics of the input video. Thus, there is no single best method for all scenarios. A combination of the above factors i.e. hybrid models significantly influence the overall design choices made at each stage within the summarization pipeline.

IV.1 LESSONS LEARNED FROM STAGE-WISE DESIGN CHOICES

- **Segmentation and Pre-processing:** Segmentation based on deep learning has been shown to significantly improve consistency when capturing shots and event detection, especially in heterogeneous/untrimmed video content. But rigid boundaries around segmentation can introduce early errors into the subsequent processing stages, resulting in an error propagation. Approaches that either use hybrid strategies that soften the segmentation decision or incorporate detection of boundaries into temporal modeling provide much better error resilience. Therefore, segmentation should be viewed as a semantic abstraction problem instead of a hard preprocessing step.

- **Feature Representation:** The most significant factor for summary quality is learned multimodal representation, especially when combining visual, audio, and linguistic cues. Vision-Language Models and transformer-based encoders demonstrate much greater consistency in cross-dataset generalization than handcrafted feature-based approaches, as vision-language models encode high-level semantics as opposed to low-level statistical appearance characteristics. Handcrafted features can provide value due to their efficiency; however, they do not convey the narrative intent or context, indicating that the richness of the semantic content and compactness is related to producing effective summaries.

- **Temporal Modeling:** Models which are able to learn long range dependencies, like transformers and hierarchical attention mechanisms, perform well with long-form data sets like ActivityNet; however, due to their quadratic complexity they present scalability issues. The best designs are those that utilize hybrid time reasoning using local convolutional models with sparse or hierarchical attention that will provide both temporal fidelity and computationally feasibility, therefore making them practical solutions to many forms of video summarization.

IV.2 RELEVANCE SCORING, SELECTION, AND DATASET AWARENESS

The significance of relevance in a video summary is determined by a series of decisions that are based upon preference rather than rated score values. Although many supervised models have succeeded in meeting their benchmarks and perform well according to these metrics, they still get affected by annotator bias, and generalization. Self-supervised and reinforcement-based methods are better suited for adapting by focusing on diversity, representativeness, and temporal consistency, than supervised methods. The best performing summary systems use a combination of heuristic diversity rules along with attention strategies and reward optimization to heavily balance video summaries while reducing the potential for overfitting general to the annotators. Additionally, dataset analysis illustrates how success on short-form benchmarks such as SumMe and TVSum do not guarantee robustness when applied to real-world problems. Models trained and/or tested using long-form, untruncated datasets demonstrate significantly greater scalability and overall reliability, indicating that dataset diversity and duration sensitivity should be treated as a fundamental building block in designing an effective video summarization system rather than merely bilateral validation tools.

IV.3 COMPUTATIONAL TRADE-OFFS AND GENERALIZATION INSIGHTS

Light, efficient models are ideal for real-time use and at the edge of networks, offering the best combination of speed and low latency with good accuracy comparable to high end models. While there are powerful transformer models to perform complex analysis in offline environments and reasoning with depth, there is diminishing accuracy for the time used to compute versus the quality of the output. Alternatives to both heavy transformers or light transformers that still maintain sufficient depth are optimized transformers and more generalized use of transformers with strategies such as knowledge distillation or token sparsification. Models that have high generalization and robustness to domain shifts (shifting from one domain to another) have used multimodal learning, hierarchical temporal abstraction and model uncertainty (being aware of the stability of the model) to help reduce the effect of domain shifts, differences in video length, and differing video content. Also, pre-trained models that either have no training or weakly supervised learning frequently achieve significant stability by avoiding overfitting to datasets (making sure there aren't too many examples), which demonstrates the key role of using large pre-trained representation of models.

IV.4 KEY DESIGN INSIGHTS AND BEST PRACTICES

The insight obtained from the analysis are as follows,

- Both hybrid model architectures produce the best overall results by combining the optimization of representation learning, temporal reasoning, and adaptive selection together.
- The most scalable model architectures produce highly efficient scalable output at a low computational cost.
- The models with the greatest degree of generalization are video architectures trained on diverse, long form datasets (Vision & Local Models and Multimodal Models). Hence these models can generalize to many different domains of video data.
- Combining soft-segmentation, preference aligned relevance scoring and adaptive post-processing outputs is among the strongest video summarization methods with an end-to-end strategy creating reliable, balanced outputs.

Effective video summarization analysis shows that to achieve an effective output for the final video summary; one centralized design needs to be developed instead of optimizing standalone pipelines. Future video summarization models must place greater emphasis on semantic alignment; which need a dataset aware method of measuring performance. Finally, temporal modelling methods must use efficiently as their basis for producing consistent performance results.

V. ROADMAP AND FUTURE RESEARCH DIRECTIONS

Video summarization has entered a new, critical step of evolution, characterized by large-scale (increasing amounts of video) and multimodal (adding different types of audio/video streams) complexity as well as deployment diversity. Although, deep learning models have made a significant contribution to improving summarization quality, continuing challenges associated with scaling, generalizing, interpreting, and evaluating highlight the need for a structured research roadmap to address these issues. This section will provide an overview of the significant future directions for developing video summarization systems.

V.1 SCALABLE LONG-VIDEO AND STREAMING SUMMARIZATION

The focus of future research needs to be on creating long-term summaries for continuous video streams in real-world situations with limited memory and latency. To process lengthy video, some methods need to be employed, which include hierarchical temporal abstractions, sparse and adaptive attention mechanisms, Memory-Aided Transformers and Streaming-Aware Architectures. Another major area of research is the development of incremental and online summarization techniques to support real-time summary creation and continuous update of summaries with global narrative coherence for use in surveillance systems, autonomous systems, and live content platforms.

V.2 PREFERENCE-AWARE, PERSONALIZED, AND EXPLAINABLE SUMMARIZATION

Most of the currently available summaries focus on generic ranking metrics but do not consider the user's intent and context. In the future, new summarization systems will need to incorporate user preference-based learning methods. These methods will allow for more personalized summaries based on the user's profile and goals as well as the semantics of their queries. In addition, incorporating "human-in-the-loop" (i.e., human feedback) within reinforcement learning for preference optimization will help facilitate adaptive relevance estimation.

In addition, additional methods for increasing interpretability and explainability (e.g., attention visualization, semantic reasoning graphs, causal analysis) will create confidence, transparency and usability in high-stakes decision-making environments.

V.3 ROBUST EVALUATION PROTOCOLS AND DATASET EVOLUTION

Progress based on benchmarks is limited due to issues such as annotation subjectivity, bias within datasets and low temporal diversity. For future work on large-scale datasets that accurately represent a wide range of real-world videos, a focus on developing diverse datasets with weak supervision will be necessary. Metrics used in assessing systems should include semantic coverage, temporal consistency, user satisfaction, as well as computational efficiency and should not focus solely on F-score. Cross-dataset and zero-shot evaluation should be standard ways to assess generalization and robustness across systems, so they can be effectively evaluated to determine performance on new datasets.

V.4 EFFICIENCY-AWARE AND RESOURCE-CONSTRAINED DEPLOYMENT

To allow for the larger scale use of video summarization techniques and methods, there is a need to create new models that consider potential deployment issues as well as the need for efficiency. Efficiency-aware video summarization models need to be created using video summarization model compression, knowledge distillation, and quantization to allow for their efficient deployment on edge and mobile devices. In addition, hybrid cloud-edge deployments and inference methods will be required to support these models when they are deployed into production. The future directions for video summarization will involve integrating scalable temporal reasoning, preference-aligned learning, and evaluation. This path will provide the framework for the video summarization field to transition from optimizing to benchmark-based approaches in creating scalable and interpretable video summarization systems that will address many issues with understanding large amounts of video content.

VI. CONCLUSION

This review provides a detailed overview of the progression of video summarizing techniques from heuristic methods through ML and DL to hybrid techniques across segmentation, feature extraction, temporal modelling, relevance scoring, and post processing. Comparative evaluations of DL versus hybrid architectures using standard benchmark datasets, indicate significant improvements regarding semantic representation, temporal coherence and robustness. Moreover, the evaluations highlight numerous challenges associated with these approaches such as annotation bias, error propagation across modules, computational complexity, and a lack of generalizability for long-form or diverse video content. Furthermore, the results of a data-centric assessment utilizing complexity-aware analytical techniques have shown that high benchmark scores do not necessarily correlate with true effectiveness in real-world applications, revealing a need for more diverse, scalable, and efficient design criteria that take into account the size and characteristics of their respective sample populations. Based on the findings from this review, there is a need for a significant change in the industry for a unified multi-modal and preference-based framework to achieve a good balance of semantic richness and practical deployment issues. Future scope of this research is to develop a complete architecture that allows for scalable long-term reasoning, as well as a system that offers personalized preference-based deployment for video summarization solutions. Moreover, it is important to create solid evaluation protocols and design performance-optimized models to provide video summarization systems with the reliability needed to run properly across a wide variety of applications, thereby allowing these solutions to operate successfully in a variety of real-world environments.

VII. AUTHOR'S CONTRIBUTION

Conceptualization: Aruna M, Rajiv Ranjan Singh.

Methodology: Sagar Bhutada.

Investigation: Sagar Bhutada.

Discussion of results: Sagar Bhutada.

Writing – Original Draft: Sagar Bhutada.

Writing – Review and Editing: Sagar Bhutada.

Resources: Sagar Bhutada.

Supervision: Sagar Bhutada.

Approval of the final text: Sagar Bhutada.

VIII. REFERENCES

- [1] N. Willis, A. Bernstein, and L. Rossetto, "Effects of human cognition-inspired task presentation on interactive video retrieval", *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 21, no. 6, pp. 1–25, 2025.
- [2] D. Vora, P. Kadam, D. D. Mohite, N. Kumar, N. Kumar, P. Radhakrishnan, and S. Bhagwat, "AI-driven video summarization for optimizing content retrieval and management through deep learning techniques", *Sci. Rep.*, vol. 15, no. 1, p. 4058, 2025.
- [3] P. Song, L. Zhang, L. Lan, W. Chen, D. Guo, X. Yang, and M. Wang, "Towards efficient partially relevant video retrieval with active moment discovering", *IEEE Trans. Multimedia*, vol. 27, pp. 6740–6751, 2025.
- [4] J. Paul, A. Roy, A. Mitra, and J. Sil, "HyV-Summ: Social media video summarization on custom dataset using hybrid techniques", *Neurocomputing*, vol. 614, p. 128852, 2024.
- [5] C. S and A. Kumar, "Contextual feature fusion-based keyframe selection using semantic attention and diversity-aware optimization for video summarization", *Symmetry*, vol. 17, no. 10, p. 1737, 2025.

- [6] E. Marevac, E. Kadušić, N. Živić, N. Buzadžić, E. Tabak, and S. Velić, “Multimodal video summarization using machine learning: A comprehensive benchmark of feature selection and classifier performance”, *Algorithms*, vol. 18, no. 9, p. 572, 2025.
- [7] S. B. Veesam and A. R. Satish, “An empirical taxonomy of video summarization model from a statistical perspective”, *IEEE Access*, vol. 12, pp. 173850–173866, 2024.
- [8] Y. Zhang, Y. Liu, W. Kang, and R. Tao, “VSS-NET: Visual semantic self-mining network for video summarization”, *IEEE Trans. Circuits Syst. Video Technol.*, vol. 34, no. 4, pp. 2775–2788, 2023.
- [9] Y. Jin, X. Tian, Z. Zhang, P. Liu, and X. Tang, “C2F: An effective coarse-to-fine network for video summarization”, *Image Vis. Comput.*, vol. 144, p. 104962, 2024.
- [10] J. Xie, X. Chen, S. Zhao, and S. Lu, “Video summarization via knowledge-aware multimodal deep networks”, *Knowl.-Based Syst.*, vol. 293, p. 111670, 2024.
- [11] Y. Zhang, Y. Liu, and C. Wu, “Attention-guided multi-granularity fusion model for video summarization”, *Expert Syst. Appl.*, vol. 249, p. 123568, 2024.
- [12] J. Qin, H. Yu, W. Liang, and D. Ding, “Video summarization using knowledge distillation-based attentive network”, *Cogn. Comput.*, vol. 16, no. 3, pp. 1022–1031, 2024.
- [13] L. Yu, X. Zhao, L. Xie, H. Liang, and R. Liang, “Hierarchical multimodal video summarization with dynamic sampling”, *IET Image Process.*, vol. 18, no. 14, pp. 4577–4588, 2024.
- [14] J. Jin and S. A. Sharudin, “GARNN-AE-LSTM: A multimodal deep learning approach for high-accuracy video summarization”, *J. Vis. Exp.*, no. 224, 2025.
- [15] V. Q. Bang and V. H. Viet, “Integrate the temporal scheme for unsupervised video summarization via attention mechanism”, *IEEE Access*, vol. 13, pp. 38147–38162, 2025.
- [16] X. Min, H. Duan, W. Sun, Y. Zhu, and G. Zhai, “Perceptual video quality assessment: A survey”, *Sci. China Inf. Sci.*, vol. 67, no. 11, 2024.
- [17] S. C. Hadke and R. Mishra, “Shot boundary detection in video using dual-stage optimized VGGNet based feature fusion and classification”, *Multimed. Tools Appl.*, vol. 83, no. 12, pp. 34417–34444, 2023.
- [18] A. Benoughidene, F. Titouna, and A. Boughida, “Static video summarization based on genetic algorithm and deep learning approach”, *Multimed. Tools Appl.*, vol. 84, no. 13, pp. 12487–12512, 2024.
- [19] P. C. Kaur and L. Ragha, “Optimized deep learning enabled lecture audio video summarization”, *J. Vis. Commun. Image Represent.*, vol. 104, p. 104309, 2024.
- [20] P. Singh and A. K. S. Kushwaha, “Key frame extraction algorithm for video summarization based on sliding window”, *Multimed. Tools Appl.*, vol. 84, no. 26, pp. 31793–31812, 2024.
- [21] O. Issa and T. Shanableh, “Static video summarization using video coding features with frame-level temporal subsampling and deep learning”, *Appl. Sci.*, vol. 13, no. 10, p. 6065, 2023.
- [22] M. Su, R. Ma, B. Zhang, and K. Li, “Recurrent unit augmented memory network for video summarisation”, *IET Comput. Vis.*, vol. 17, no. 6, pp. 710–721, 2023.
- [23] K. Zhao, R. Ma, M. Su, H. Yu, P. An, and G. Li, “MOFNET: Memory optimization fusion network for query-focused video summarization”, *IEEE Sens. J.*, vol. 25, no. 14, pp. 27642–27652, 2025.
- [24] Y. Wang, H. Zhang, Y. Yue, S. Song, C. Deng, J. Feng, and G. Huang, “UNI-ADAFoCus: Spatial-temporal dynamic computation for video recognition”, *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 47, no. 3, pp. 1782–1799, 2024.
- [25] K. Yashwanth and B. Soni, “Encoder-decoder architectures based video summarization using key-shot selection model”, *Multimed. Tools Appl.*, vol. 83, no. 11, pp. 31395–31415, 2023.
- [26] S. Das, S. Ghosh, A. Tiwari, R. E. Z. M. Lynghoi, S. Saha, Z. Murad, and A. Maurya, “Presenting an order-aware multimodal fusion framework for financial advisory summarization with an exclusive video dataset”, *IEEE Access*, vol. 13, pp. 48367–48379, 2025.
- [27] X. Wu, M. Ma, S. Wan, X. Han, and S. Mei, “Multi-scale deep feature fusion based sparse dictionary selection for video summarization”, *Signal Process. Image Commun.*, vol. 118, p. 117006, 2023.
- [28] A. Sabha and A. Selwal, “Data-driven enabled approaches for criteria-based video summarization: A comprehensive survey, taxonomy, and future directions”, *Multimed. Tools Appl.*, vol. 82, no. 21, pp. 32635–32709, 2023.
- [29] B. Mohamed and B. A. Yassine, “Enhanced video temporal segmentation using a Siamese network with multimodal features”, *Signal Image Video Process.*, vol. 17, no. 8, pp. 4295–4303, 2023.
- [30] M. Mao, H. Va, A. Lee, and M. Hong, “Supervised video cloth simulation: Exploring softness and stiffness variations on fabric types using deep learning”, *Appl. Sci.*, vol. 13, no. 17, p. 9505, 2023.
- [31] K. A. Chaoui, H. E. Fadil, O. Choukai, and O. A. Omar, “A wavelet–attention–convolution hybrid deep learning model for accurate short-term photovoltaic power forecasting”, *Forecasting*, vol. 7, no. 3, p. 45, 2025.
- [32] W. Lee, Y. Shi, H. Yu, L. Cheng, X. Wang, Z. Yan, and F. Kong, “HPFoRMer: Low-parameter transformer with temporal dependency hierarchical propagation for health informatics”, *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 47, no. 11, pp. 10770–10786, 2025.
- [33] G. Rafiq, M. Rafiq, and G. S. Choi, “Video description: A comprehensive survey of deep learning approaches”, *Artif. Intell. Rev.*, vol. 56, no. 11, pp. 13293–13372, 2023.
- [34] Z. Pang, Y. Nakashima, M. Otani, and H. Nagahara, “Unleashing the power of contrastive learning for zero-shot video summarization”, *J. Imaging*, vol. 10, no. 9, p. 229, 2024.
- [35] H. Khan, T. Hussain, S. U. Khan, Z. A. Khan, and S. W. Baik, “Deep multi-scale pyramidal features network for supervised video summarization”, *Expert Syst. Appl.*, vol. 237, p. 121288, 2023.

- [36] Y. Yuan and J. Zhang, "Unsupervised video summarization via deep reinforcement learning with shot-level semantics", *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 1, pp. 445–456, 2022.
- [37] K. Jin, H. Li, H. Li, Q. Liu, R. Chen, and S. Guo, "An unsupervised video summarisation method based on temporal convolutional networks", *Int. J. Comput. Appl. Technol.*, vol. 76, no. 1–2, pp. 1–16, 2025.
- [38] R. L. Testa, A. Machado-Lima, and F. L. S. Nunes, "Enhancing temporal coherence in image-to-video facial expression synthesis: A dual-loss framework for smoother generation", *IEEE Access*, 2025.
- [39] S. Kaur, L. Kaur, and M. Lal, "An effective key frame extraction technique based on feature fusion and fuzzy-C means clustering with artificial hummingbird", *Sci. Rep.*, vol. 14, no. 1, p. 26651, 2024.
- [40] D. M. Davids, A. A. E. Raj, and C. S. Christopher, "Hybrid multi-scale hard switch YOLOv4 network for cricket video summarization", *Wireless Netw.*, vol. 30, no. 1, pp. 17–35, 2023.
- [41] Z. Hu, A. Sun, Z. Wang, J. Li, Z. Song, R. Hong, and M. Wang, "Adaptive dual video summarization: From dynamic keyframes to captions", *IEEE Trans. Multimedia*, vol. 27, pp. 7040–7052, 2025.
- [42] J. Gupta, D. Rai, and P. Singh, "GenSumNet: A genre-specific and context-aware video summarization framework with adaptive thresholding and hybrid temporal modeling", *Expert Syst. Appl.*, vol. 279, p. 127360, 2025.
- [43] P. Saini, K. Kumar, S. Kashid, A. Saini, and A. Negi, "Video summarization using deep learning techniques: A detailed analysis and investigation", *Artif. Intell. Rev.*, vol. 56, no. 11, pp. 12347–12385, 2023.
- [44] U. N. Yoon, M. D. Hong, and G. Jo, "Unsupervised video summarization based on deep reinforcement learning with interpolation", *Sensors*, vol. 23, no. 7, p. 3384, 2023.
- [45] O. Issa and T. Shanableh, "Static video summarization using video coding features with frame-level temporal subsampling and deep learning", *Appl. Sci.*, vol. 13, no. 10, p. 6065, 2023.
- [46] S. F. Ahmed, M. S. B. Alam, M. Hassan, M. R. Rozbu, T. Ishtiaq, N. Rafa, M. Mofijur, A. B. M. S. Ali, and A. H. Gandomi, "Deep learning modelling techniques: Current progress, applications, advantages, and challenges", *Artif. Intell. Rev.*, vol. 56, no. 11, pp. 13521–13617, 2023.
- [47] Z. Zhang, D. Xu, W. Ouyang, and L. Zhou, "Dense video captioning using graph-based sentence summarization", *IEEE Trans. Multimedia*, vol. 23, pp. 1799–1810, 2025.
- [48] W. Li, W. Han, L. Deng, R. Xiong, and X. Fan, "Spiking variational graph representation inference for video summarization", *IEEE Trans. Image Process.*, vol. 34, pp. 5697–5709, 2025.
- [49] H. B. U. Haq, W. Suwansantisuk, and K. Chamnongthai, "Effective multi-object of interest-based video summarization using deep learning: A smart technique for user's desired content optimization", *IEEE Access*, vol. 13, pp. 208563–208589, 2025.
- [50] T. Han, Q. Zhou, J. Yu, Z. Yu, J. Zhang, and S. Zhao, "Effective video summarization by extracting parameter-free motion attention", *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 20, no. 7, pp. 1–20, 2024.
- [51] H. Liu, M. Galindo, H. Xie, L. Wong, H. Shuai, Y. Li, and W. Cheng, "Lightweight deep learning for resource-constrained environments: A survey", *ACM Comput. Surv.*, vol. 56, no. 10, pp. 1–42, 2024.
- [52] X. Sheng, A. Yu, M. Zhang, G. An, J. Park, and K. Cho, "Integrating temporal event prediction and large language models for automatic commentary generation in video games", *Mathematics*, vol. 13, no. 17, p. 2738, 2025.
- [53] S. Muksimova, S. Umirzakova, M. Sultanov, and Y. I. Cho, "Cross-modal transformer-based streaming dense video captioning with neural ODE temporal localization", *Sensors*, vol. 25, no. 3, p. 707, 2025.
- [54] W. Li, M. Liu, H. Liu, P. Wang, S. Lu, and N. Sebe, "H2OT: Hierarchical hourglass tokenizer for efficient video pose transformers", *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 48, no. 1, pp. 512–526, 2025.
- [55] C. M. Thwal, M. N. Nguyen, Y. L. Tun, S. T. Kim, M. T. Thai, and C. S. Hong, "OnDev-LCT: On-device lightweight convolutional transformers towards federated learning", *Neural Netw.*, vol. 170, pp. 635–649, 2023.
- [56] J. Lin, H. Hua, M. Chen, Y. Li, J. Hsiao, C. Ho, and J. Luo, "VideoXum: Cross-modal visual and textual summarization of videos", *IEEE Trans. Multimedia*, vol. 26, pp. 5548–5560, 2023.
- [57] N. X. Gu, "Semantic-conditional network for micro-video summarization", *Comput. Fraud Secur.*, pp. 80–91, 2025.
- [58] P. Kadam, D. Vora, S. Patil, S. Mishra, and V. Khairnar, "Behavioral profiling for adaptive video summarization: From generalization to personalization", *MethodsX*, vol. 13, p. 102780, 2024.