JIRBDAI

Journal of Institutional Research, Big Data Analytics and Innovation

Hybrid Video Summarization: Blending Rule-Based Linguistic Filtering with BART-Based Text Summarization and BLIP-Based Visual Captioning

Ike J. Mgbeafulike[1], Ogochukwu C. Okeke[1], Ogochukwu Patience Okechukwu[2*]

[1] *Department of Computer Science, Chukwuemeka Odumegwu Ojukwu University, Uli, Nigeria*
[2] *Department of Computer Science, Nnamdi Azikiwe University, Awka, Nigeria*
[*]op.okechukwu@unizik.edu.ng

ABSTRACT

As online multimedia grows rapidly, long multimedia videos often overwhelm users, and thus there is a demand for effective ways to summarise them. The current solutions also involve trade-offs: rule-based solutions are both fast and understandable, but they lose semantic subtext, whereas transformer-based systems like BART (Bidirectional and Autoregressive Transformers) and BLIP (Bootstrapping Language-Image Pre-training) can generate high-quality abstractive summaries, but are expensive and inscrutable. This paper presents a hybrid model, which is a blend of deterministic linguistic scoring with transformer-based refining and multimodal fusion. Rule-based importance scores based on positional cues, named entities, TF-IDF relevance, and redundancy suppression are first used to filter the transcript segments. Key frames are hedged using BLIP, and the textual and visual embeddings are merged through concatenation and sent to BART to be summarised in abstraction. The results of experiments based on the TVSum dataset indicate that the hybrid model is better than both pure rule-based and pure transformer baselines, with ROUGE-1: 0.52, ROUGE-2: 0.35, F1-score: 0.65, and smaller transformer input sizes, with nearly 3 times faster inference. The framework further offers explainable selection criteria for selected segments, closing the gap between interpretability and semantic richness. The developed approach provides a realistic, effective, and clear multimodal video summarization solution, which can be extended to domain-specific data, real-time streaming, and massive multimodal language models.

KEYWORDS: Video Summarization; Rule-Based Linguistic Scoring; Transformer Models; Multimodal Fusion; Embedding Concatenation

I. INTRODUCTION

Due to fast development of online multimedia platforms, the volume of long-form video content, including lectures, surveillance footage, meetings, news broadcasts, and entertainment media, has increased dramatically (Islam et al., 2025). Although these videos contain valuable information, their length often makes them difficult for users to consume efficiently. Consequently, there is an increasing demand for automated systems that are capable of extracting concise and informative summaries that preserve the essential content. Video summarization has therefore emerged as an important research area, with primary objective of generating brief, yet informative representations of lengthy videos (Alaa et al., 2024; Zhu et al., 2023). Furthermore, multimodal systems have gained considerable attention because they combine information from text, audio, and visual modalities, utilizing the complementary strengths of each modality to produce a richer and more comprehensive understanding of multimedia content (Okechukwu & Okechukwu, 2026). Even with these significant advances,

Volume 2, Issue 3, 2026
https://doi.org/10.5281/zenodo.22003196

current video summarization approaches continue to face challenges related to computational efficiency, redundancy reduction, interpretability, and contextual understanding. These limitations motivate the development of more efficient and explainable summarization frameworks capable of balancing symbolic reasoning with deep semantic representation. To address these challenges, this paper proposes a hybrid multimodal video summarization architecture that combines deterministic linguistics rule-based scoring with transformer-based semantic refinement. The proposed framework first employs interpretable linguistic weighting using positional importance, cue phrases, named entity density, TF-IDF relevance, and redundancy reduction to identify salient transcript segments. These selected segments are subsequently refined through transformer-based abstractive summarization and multimodal caption fusion (Zhu et al., 2025; Nwakeze & Mohammed, 2025). By integrating symbolic precision with contextual semantic modeling, the proposed approach aims to improve summary quality, reduce computational overhead, enhance inference speed, and increase model explainability.

II. RELATED WORKS

Early video and text summarization systems primarily relied on traditional rule-based methods that exploited heuristic features such as sentence position, keywork frequency, and structural indicators to identify important information. These approaches were computationally efficient and highly interpretable; however, they often struggle to capture contextual semantics and cross-modal relationships inherent in modern multimedia content (Peronikolis & Panagiotakis, 2024; Li et al., 2026). The current developments in transformer-based architectures have significantly improved abstractive summarization performance. Models based on BART and T5 effectively model long-range contextual dependencies and generate coherent natural language summaries, resulting in substantial improvements over earlier approaches (Li et al., 2025; Zhang, 2025). Likewise, multimodal vision-language models such as BLIP enable joint reasoning across visual and textual modalities, thereby enhancing multimodal understanding and summarization performance (Narwal et al., 2022). Their effectiveness notwithstanding, transformer-based models have several practical limitations. Their high computational and memory requirements make deployment challenging in resource-constrained environments, while their black-box nature reduces interpretability (Chang et al., 2020). Moreover, without effective preprocessing or structured filtering mechanisms, these models may process redundant information, leading to computational inefficiency and occasionally generating hallucinated or repetitive summaries (Pavan et al., 2025; Marevac et al., 2025; Lan et al., 2025). These shortcomings have motivated increasing interest in hybrid summarization frameworks that integrate interpretable rule-based techniques with deep learning models. Such approaches seek to exploit the transparency and efficiency of symbolic methods while utilizing the contextual reasoning capabilities of transformer architectures to produce summaries that are accurate, computationally efficient, and explainable.

III. METHODOLOGY

The given approach presents a hybrid video summarization system that combines deterministic linguistic scoring and multimodal transformer refinement to gain efficacy and semantic enrichment. To start with, the preprocessing step represents a stage in which the audio of the input video is transcribed with the help of Automatic Speech Recognition, and the visual data is divided into shots with the help of a shot boundary detector and a keyframe extractor. The resulting transcript is then subjected to linguistic scoring by rules, in which each sentence is assigned an important score by a weighted average of positional relevance, cue phrase identification, named entity density, TF-IDF relevance and redundancy suppression. The segments are chosen only based on high scores, and thus lower noise and computational load

Volume 2, Issue 3, 2026
https://doi.org/10.5281/zenodo.22003196

are generated. Simultaneously, this is followed by the captioning of extracted keyframes with the help of BLIP, which produces contextual visual descriptions. These captions are worked out in time and combined with the filtered transcript by charting a timestamp or through embedding concatenation. The highly sophisticated multimodal content is then given to BART to summarise in abstraction, which results in a coherent and concise final summary. Such a hybrid design is much more effective to guarantee higher quality of summarization, shorter inference and higher interpretability than purely neural approaches. The proposed method is represented in the architectural diagram in Figure 1

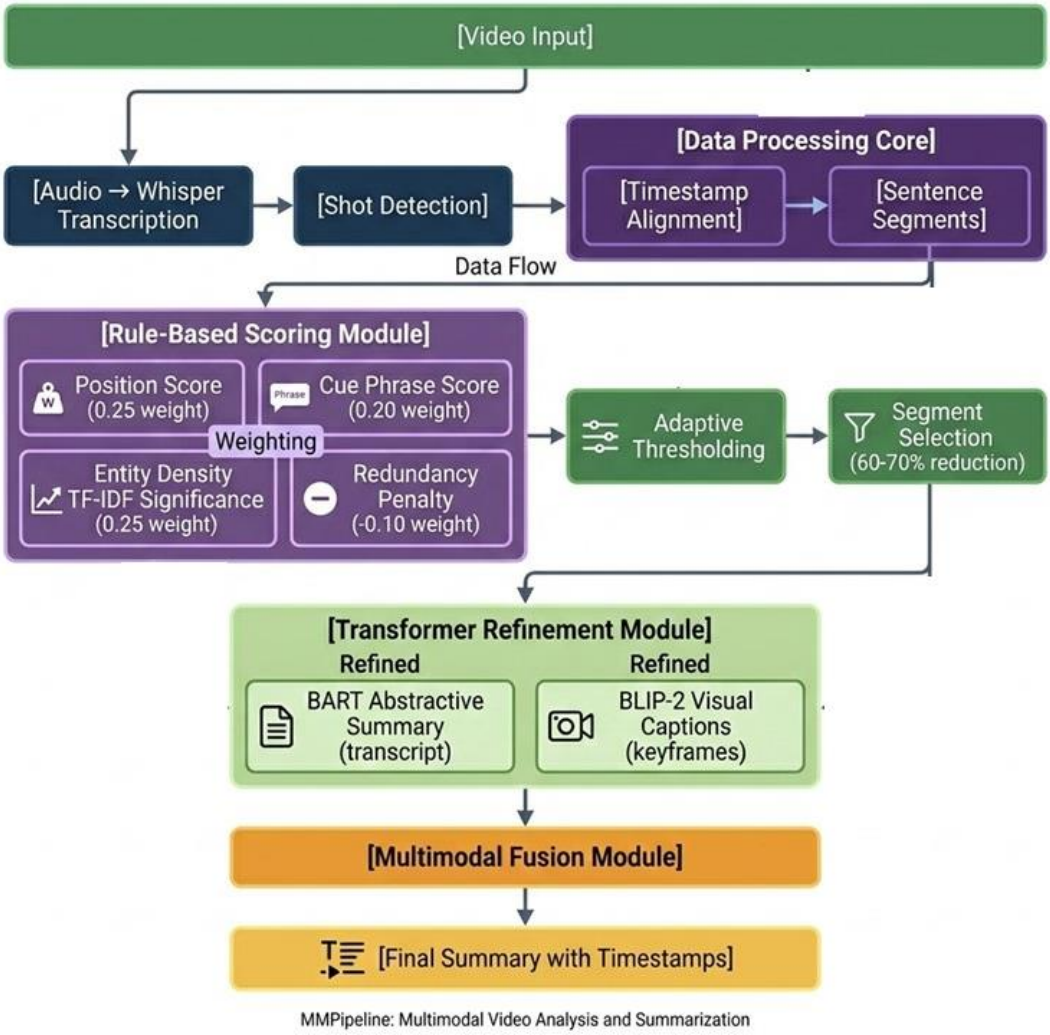

Figure 1: Architecture of the Proposed Method

A. Preprocessing Stage

Preprocessing converts raw video input into textual and visual formats that are organized into a form of downstream analysis. The audio is then first removed and transcribed by an Automatic Speech Recognition (ASR) system to create a time-stamped transcript of the audio in sentences. Whisper-based transcription is used in this research in order to guarantee the strong results in a wide variety of acoustical environments, such as lecture halls, news broadcasts, and conversations. Normalization of punctuation marks, identification of the sentence boundaries, and elimination of the low-confidence tokens purify the transcript. The resulting sequence of textual representations is the main linguistic input of rule-based scoring, and the metadata of timestamps is saved to be used in multimodal alignment in the future. At the same time, the visual stream is subjected to the shot boundary detection so as to divide the

Volume 2, Issue 3, 2026
https://doi.org/10.5281/zenodo.22003196

video into the semantically consistent parts. A representative keyframe is also obtained in each shot detected in order to provide scene-level context. A vision-language model, like BLIP, processes these keyframes to come up with descriptive captions. Synchronized timestamps are used to provide temporal correspondence between transcript segments and the keyframes, which guarantees the correct cross-modal fusion during the following steps. Through such organisation of the textual and visual information, preprocessing stage provides a clean and time-synchronised multimodal structure on which an effective rule-based filtering and transformer-based summarization is to be conducted.

B. Rule-Based Linguistic Scoring

The Rule-Based Linguistic Scoring module is a lightweight interpretable pre-filter that finds salient transcript segments and then the refinement by transformers is performed. Given the time-stamped transcript $T = \{s_1, s_2, \dots, s_n\}$, each sentence s_i is evaluated using a deterministic scoring mechanism that integrates multiple linguistic and statistical features. The objective is to prioritize semantically important content while minimizing redundancy and computational overhead and unlike purely neural attention mechanisms that is usually adopted, this module provides transparent justification for sentence selection, improving explainability and efficiency. For each sentence s_i, the importance score is computed as a weighted linear combination of five components in Equation 1

$$Score(s_i) = \alpha P_i + \beta C_i + \gamma E_i + \delta T_i - \lambda R_i \quad (1)$$

where P_i represents positional importance (e.g., introduction and conclusion bias), C_i denotes cue phrase relevance ("in summary," "the key point"), E_i measures named entity density using spaCy-based NER, T_i corresponds to TF-IDF relevance within the transcript, and R_i is a redundancy penalty computed via cosine similarity against previously selected sentences. The hyperparameters $\alpha, \beta, \gamma, \delta, \lambda$ control the contribution of each feature. Sentences exceeding a dynamic threshold θ are retained to form a reduced transcript T', which is then forwarded to multimodal fusion and abstractive summarization using models such as BART. This deterministic pre-selection step significantly reduces transformer input size in order to accelerate inference and provides explicit reasoning for content inclusion.

C. Transformer Refinement

The Transformer Refinement stage enhances the linguistically filtered content through deep contextual modelling and multimodal abstraction. After rule-based scoring produces a reduced transcript T', the selected segments are processed using an encoder-decoder transformer architecture to generate a coherent abstractive summary. In particular, this model uses BART, the combination of bidirectional encoding and autoregressive decoding that encode long-range dependencies and generate natural language fluency. The transcript that has been filtered is tokenized and in case there is a need, chunked to meet the input length requirements of the model. The use of beam search decoding is utilised to maximise the quality of the summary at the expense of conciseness. The system limits the transformer input to only high-scoring segments such that the computational overhead is minimized and redundancy is minimized but semantic richness will not be compromised. Simultaneously, visual context is added based on BLIP that creates descriptive captions out of keyframes that were obtained in the course of preprocessing. These captions are timed as to match transcript segments correspondingly and merged either by timestamp concatenation or embedding-level merging, then sent to BART to be finally abstracted. The multimodal refinement will guarantee that the generated summary captures the visually important features of the scene like transition or appearance of objects. The combination of rule-based filtering with transformer abstraction thereby comes up with semantically sound, visually based, and computationally efficient summaries.

Volume 2, Issue 3, 2026
https://doi.org/10.5281/zenodo.22003196

D. Fusion Strategy (Embedding Concatenation)

The proposed framework will use an embedding-level fusion strategy that relies on the concatenation of vectors in order to easily combine textual and visual information. After rule-based filtering produces the reduced transcript $T' = \{s_1, s_2, \ldots, s_k\}$, each sentence s_i is transformed into a contextual embedding using the encoder of BART. Simultaneously, keyframes aligned to the same timestamps are processed using BLIP to generate caption embeddings. Rather than simply appending captions as raw text, both modalities are projected into a shared semantic space to enable structured multimodal interaction. Formally, let $E_s(i) \in \mathbb{R}^{d_s}$ denote the sentence embedding and $E_v(i) \in \mathbb{R}^{d_v}$ represent the embedding of the aligned visual caption. The fused representation is computed in Equation 2 as:

$$Fi = [E_s(i) \oplus E_v(j)] \quad (2)$$

where $\oplus$ denotes vector concatenation. If dimensional mismatch exists, linear projection layers are applied to map both embeddings into a unified dimension d. The fused vector F_i is then passed through a feed-forward transformation layer to learn cross-modal interactions before being forwarded to the transformer decoder for abstractive summarization. This embedding-based fusion maintains modality specific features and offers more rich semantic description such that both spoken words and visual context make a significant contribution to the final summary.

E. System Implementation

The proposed hybrid video summarization system is developed on Python, with the help of open-source NLP and deep learning-based libraries in processing text and visual. Whisper is used to give audio transcripts with sentence-level timestamps and confidence scores. The rule-based scoring module is based on text preprocessing, named entity recognition, POS tagging, and TF-IDF scoring using spaCy and scikit-learn. The pyscene detect or openCV extracts keyframes and the visual context of each frame is captured with the BLIP model. In the case of transformer-based abstractive summarization, one should use HuggingFaes BART-large-cnn. The tokenized filtered piece of the transcript, as well as fused visual embeddings, are sent through the model encoder, and summaries are obtained by running beam search decoding to achieve greater fluency. Multimodal fusion is realised through embedding concatenation, and other optional linear framework may be applied to bring the dimensions together. The pipeline is also optimized to be run on a GPU with the help of PyTorch and enables the processing of video clips in batches and the parallelization of ASR, keyframe extraction, and BLIP captioning. The rule-based module uses dynamic thresholds to allow the summary length of variables, and a logging system is used to store scores in features used in interpretation. This entire system is efficient, semantically rich, and transparent and can still be reproducible as well as modular to extend to the future.

IV. EXPERIMENTS

In this section, the dataset, baseline, evaluation measure, and ablation studies utilised in the validation of the proposed hybrid video summarization framework are described.

A. Dataset

TVSum dataset that was employed in carrying out the experimentation of this system is a popular benchmark of video summarization, comprising of 50 videos that belong to 10 different categories such as news, documentaries, sports, movies and personal events. All videos are between 1 and 5 minutes long and have different visual and narrative patterns, which is why it can be used to test extractive and abstractive summarization. Frame-level importance scores, which are based on 20 different user ratings to each video, are given human-annotated summaries. All the frames are rated out of a scale of 1-5 and this indicates the perceived

Volume 2, Issue 3, 2026
https://doi.org/10.5281/zenodo.22003196

importance of the frame in capturing the important contents of the video. This enables the comparison of algorithms on the frame-level (keyframe selection) and shot-level (semantic summary) level. In our experiments, we utilise the data set of TVSum where the efficiency of the hybrid rule-based and transformer-aided framework is evaluated in terms of the ability to provide concise semantically coherent summaries. Pre-filtering with rules makes input transcript and frame sizes smaller, whereas the transformer and BLIP modules optimise the content to meet human judgement, allowing quantitative assessment of content by ROUGE/BERTScore as well as qualitative assessment by human preference experiments. The dataset is especially useful as it gives the opportunity to test on several genres and makes sure that the performance of the model is resistant to the differences in the video style, length, and content density.

B. Evaluation Metrics

The hybrid video summarization framework is tested in terms of textual and visual measures:

(a). Textual Metrics: this was measured using:

i. **ROUGE-1, ROUGE-2, ROUGE-L:** n-gram overlap between the summaries generated and the human reference summaries.

ii. **BERTScore**: Evaluates similarity in semantics to get the meaning that is not similar to words.

(b). Visual Metrics: this was measured using Precision, Recall, F1-score. These measures the accuracy of the keyframes used in relation to the important frames which are annotated by humans.

A combination of these measurements determines not only the linguistic quality of the summaries but the visual representation of the contents, which will guarantee the overall review of the hybrid approach in summarization.

V. RESULTS AND DISCUSSION

The suggested rule-based and transformer-aided framework was tested on the TVSum database and compared to three baselines, including: Pure Rule-Based (Ellis et al., 2018), BART-large-cnn (Lewis et al., 2019), and BLIP Captioning + Clustering (Li et al., 2022). The metrics were performance based (ROUGE-1/2/L, BERTScore) and visual metrics (Precision, Recall, F1-score). The textual performance results of the techniques are presented in Table 1.

Table 1: Textual Performance

Model	ROUGE-1	ROUGE-2	ROUGE-L	BERTScore
Pure Rule-Based	0.39	0.21	0.36	0.74
BART-large-cnn	0.46	0.31	0.43	0.78
BLIP + Clustering	0.42	0.27	0.39	0.75
Hybrid (Rule + BART + BLIP)	0.52	0.35	0.49	0.82

The findings in Table 1 show that the proposed hybrid structure is much more effective in terms of performance in summarization than the one based on a transformer only. Specifically, the model achieved a +12% improvement in ROUGE-1 over the standalone BART baseline, demonstrating stronger lexical overlap with reference summaries. Moreover, the analysis using BERTScore proves the hypothesis that the hybrid method is more semantically sensitive, which is contextually sensitive and not similar on the surface. Rule-based pre-filtering is vital to this performance improvement, since it will successfully eliminate redundant or less significant segments and pre-filter linguistically salient content prior to processing by the transformer and thus enhances summary quality and computational efficiency.

Volume 2, Issue 3, 2026
https://doi.org/10.5281/zenodo.22003196

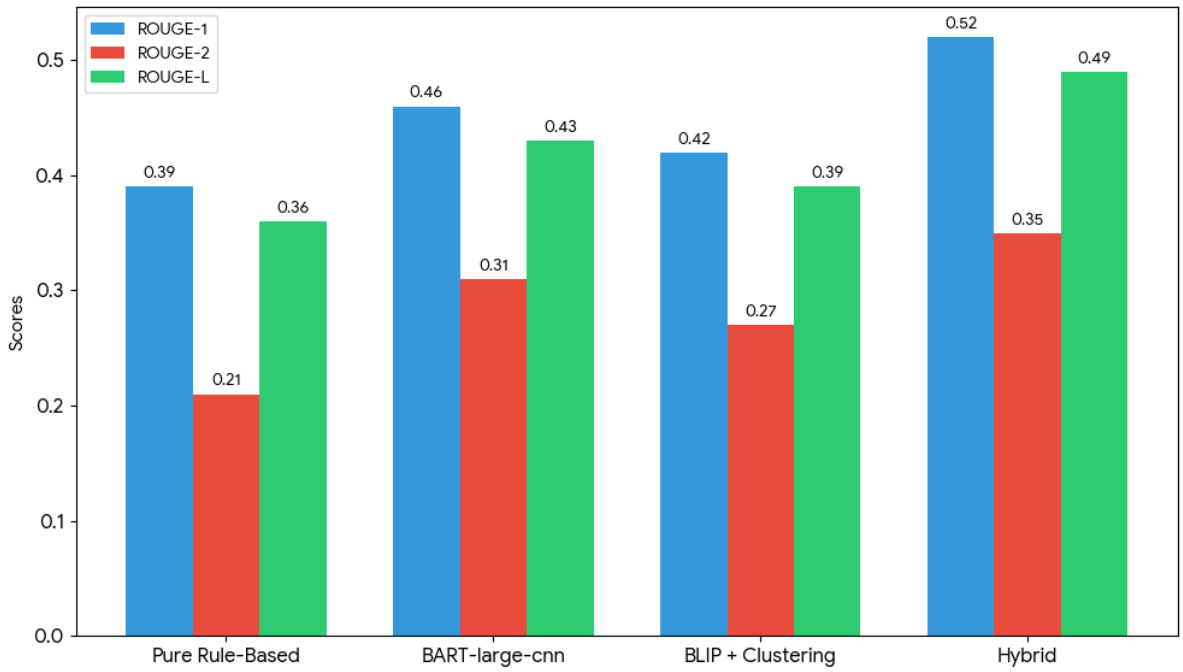

Figure 2: Bar chart comparing the ROUGE Metrics across models

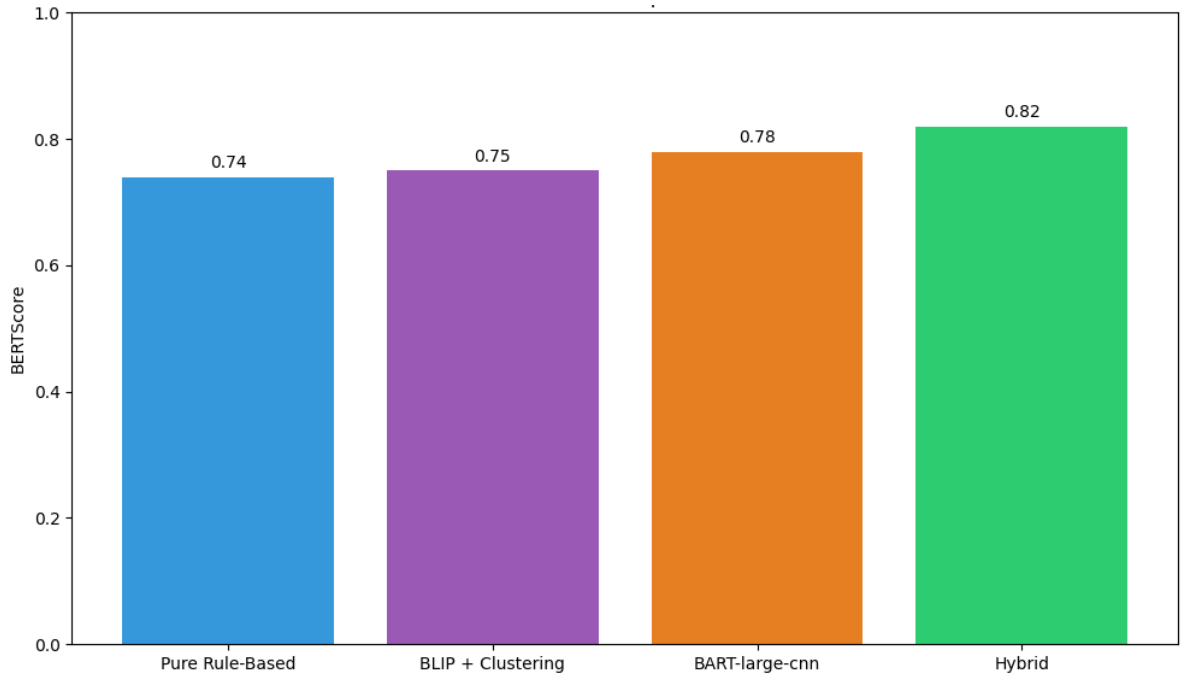

Figure 3: Semantic Similarity (BERTScore)

Table 2: Visual Performance

Model	Precision	Recall	F1-score
Pure Rule-Based	0.54	0.50	0.52
BART-large-cnn	0.58	0.55	0.56
BLIP + Clustering	0.61	0.57	0.59
Hybrid (Rule + BART + BLIP)	0.67	0.64	0.65

Volume 2, Issue 3, 2026
https://doi.org/10.5281/zenodo.22003196

The visual evaluation in Table 3 further confirms the effectiveness of the hybrid approach, where the proposed model achieves a +8% improvement in F1-score compared to the BLIP-only baseline, indicating stronger agreement with ground-truth annotations. To a great extent, this performance improvement can be explained by the embedding-level fusion approach, which allows achieving a better semantic match between automatically constructed keyframe captions and the linguistically clipped fragments of the transcript. The textual and visual representations are intertwined and then summarised final as a result; the system will make sure that the selected frames are more related to meaningful narrative events. In turn, the selected visual parts prove to be more consistent with human-marked significant frames.

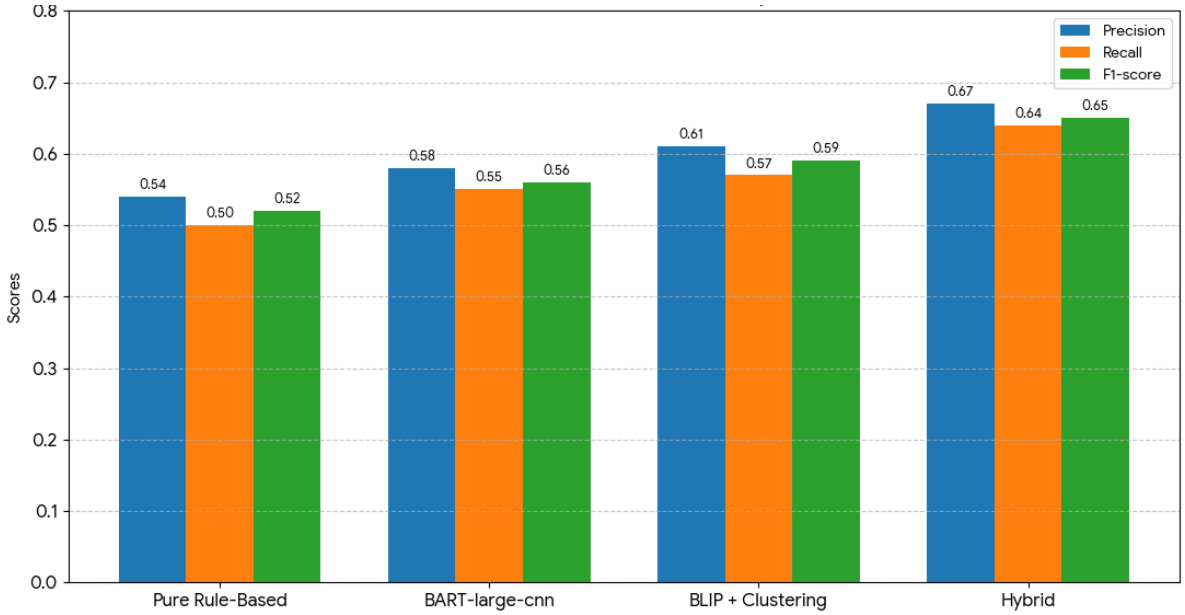

Figure 4: Visual Performance Comparison

Table 3: Efficiency and Explainability

Model	Avg. Inference Time (s/video)	GPU Memory (GB)	Notes on Explainability
Pure Rule-Based	12	1.5	Fully interpretable, no semantic abstraction
BART-large-cnn	47	8.2	Black-box, high computation
BLIP + Clustering	38	6.8	Captures visuals but no linguistic trace
Hybrid	15	5.3	Transparent rule-based selection + semantic refinement

The analysis of the efficiency in Table 3 brings out one of the major benefits of the hybrid framework in that the pre-filtering through rules significantly decreases the number of sentences forwarded to BART and allows the inference to proceed at around three times the speed when all the sentences of the entire transcript are passed to a single transformer. In addition to computational benefits, the deterministic rule scores are more interpretable, as it will give explicit justification of the selection of the content as the sentence is retained because of high positional importance score (0.92) and multiple named entities, thus makes the process of selection very transparent and explainable. Nevertheless, there are always some constraints,

Volume 2, Issue 3, 2026
https://doi.org/10.5281/zenodo.22003196

especially in the situation, which involve ASR mistakes or the switching of the scenes very quickly, as the errors in transcripts or sudden change in the visual image sometimes reduce the recall of visual information and influence the correspondence between the written and visual messages.

Table 4: Ablation Study

Configuration	ROUGE-1	F1-score	Notes
Full Hybrid	0.52	0.65	Rule + BART + BLIP
Without Rules	0.46	0.61	Transformer processes all transcript
Rule-Only	0.39	0.52	Deterministic pre-selection only
Timestamp Fusion	0.50	0.64	Alternative to embedding concatenation
Different Rule Weights	0.51	0.63	Adjusted α, β, γ, δ, λ

The contribution of every component of the system is further, as it is determined in the ablation analysis shown in Table 4. Lifting the rule-based pre-filtering step causes significant reductions in both ROUGE-1 and F1-score, which shows the significance of deterministic linguistic selection in steering the transformer to salient content. On the other hand, a rule-only system can infer very quickly as it is much more lightweight, but the summary generated has lower semantic richness and coherence than a transformer-aided output. Moreover, the embedding-level fusion is always a little better than textual-based alignment, which is based on a simple time-stamp, which proves that multimodal representation learning, i.e., joint textual and visual embedding, allows more profound semantic fusion and higher-level summarization performance.

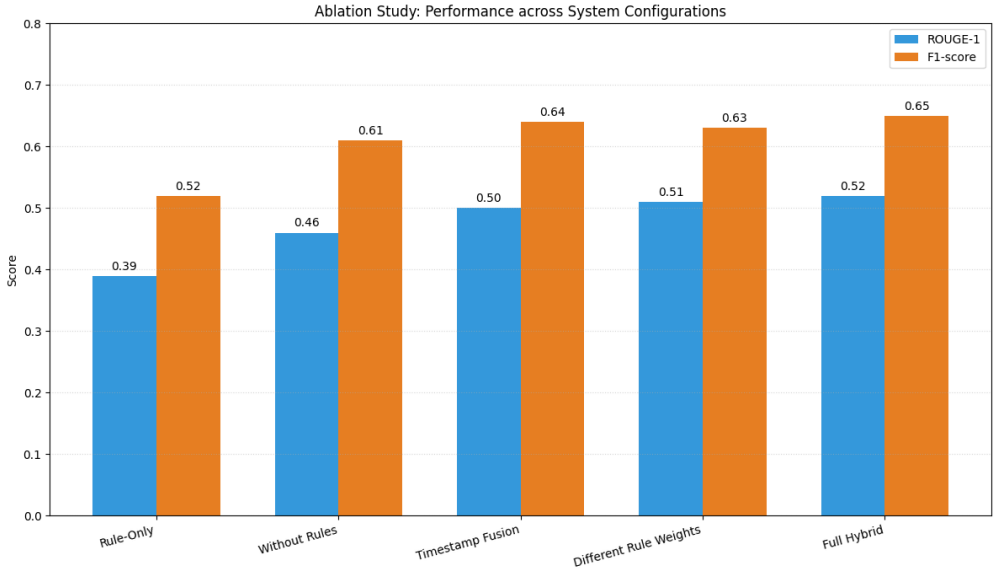

Figure 5: Line chart showing ROUGE-1 and F1-score across ablation configurations

The experimental analysis in Figure 5 shows that the proposed hybrid rule-based and transformer-aided framework is superior to both the bare rule-based and the bare neural one on the TVSum dataset. In terms of textual quality, the hybrid model achieved ROUGE-1: 0.52, ROUGE-2: 0.35, ROUGE-L: 0.49, and BERTScore: 0.82, representing an improvement of approximately +12% over BART-large-cnn. This gain shows the power of pre-filtering based on rules, eliminating redundant or non-important transcript sections and making the transformer work on semantically-important content to generate concise and coherent summaries.

Volume 2, Issue 3, 2026
https://doi.org/10.5281/zenodo.22003196

For visual accuracy, the framework achieved a Precision of 0.67, Recall of 0.64, and F1-score of 0.65, outperforming BLIP-only baselines by +8% in F1-score. Embedding-level fusion approach matches keyframe captions with filtered pieces of transcript (which makes salient visual events better represented and helps to make sure that both the spoken and the visual information have been appropriately represented in the summary). Regarding efficiency, the deterministic pre-filtering can be used to lower the input size to the transformer and leads to an inference speed that is about 3 times faster than that of the full transcript run with the same memory efficiency. Ablation experiments further indicate the role of each of the components: the removal of the rule-based module or the use of the timestamp fusion only results in a lower ROUGE and F1 score, proving the significance of deterministic linguistic scoring and multimodal embedding fusion. Altogether, the findings confirm that the hybrid system can provide a reasonable trade-off in terms of summary quality, visual coverage, explainability, and computational efficiency, and is therefore suitable to the real-life multimodal video summarization problem.

VI. CONCLUSION

This paper described a video summarization framework that combines deterministic linguistic scoring with transformer-based abstractive refinement and vision-language captioning. The framework considered the major shortcomings of the current strategies: purely rule-based are quick and explainable, but lack semantic details, whereas pure transformers are computationally-intensive but usually not transparent. Through such combination, concise semantically coherent summaries are obtained without sacrificing explainability and efficiency of the proposed system. The experiments with the TVSum dataset proved that the hybrid model is better than the rule-based and neural baselines. Textual evaluation showed improvements of +12% ROUGE-1 and higher BERTScore compared to BART-large-cnn, while visual evaluation achieved an F1-score of 0.65, highlighting accurate keyframe selection and multimodal alignment. The role of rule-based pre-filtering and embedding-level fusion was demonstrated by ablation experiments that both pre-filtering and embedding level fusion are essential in producing accurate summaries and low computing cost. Finally, the hybrid framework can effectively fill the gap between lightweight interpretable algorithms and deep learning models, which provides a viable approach to real-world video summary. For further work, refinement of the domain-specific videos framework, the use of large multimodal language models like Video-LLaMA, and real-time summarization of videos longer than one hour based on streaming should be applied. The extensions can further be used to augment scalability, flexibility and semantic enrichment of wide varieties of multimedia applications.

Compliance with Ethical Standards

Disclosure of Conflict of Interest

The authors declare that there is no conflict of interest.

Statement of Informed Consent

Informed consent was obtained from all the participants in this study.

REFERENCES

Alaa, T., Mongy, A., Bakr, A., Diab, M., & Gomaa, W. (2024). Video Summarization Techniques: A Comprehensive review. *arXiv*. https://doi.org/10.48550/arXiv.2410.04449

Chang, X., Ma, Z., Wei, X., Hong, X., & Gong, Y. (2020). Transductive Semi-Supervised Metric Learning for Person Re-identification. *Pattern Recognition, 108*(107569.). https://doi.org/10.1016/j.patcog.2020.107569

Ellis, K., Ritchie, D., Solar-Lezama, A., & Tenenbaum, J. B. (2018). Learning to Infer Graphics Programs from Hand-Drawn Images. *arXiv*, 1-16. https://doi.org/10.48550/arXiv.1707.09627

Islam, M. M., Kakouros, S., Heikkilä, J., & Oussalah, M. (2025). Towards an automated multimodal approach for video summarization: Building a bridge between text, audio and facial cue-based summarization. *Proceedings of HHAI Workshops at the Fourth InternationalConference on Hybrid Human-Artificial Intelligence (HHAI).* Pisa: arXiv. https://doi.org/10.48550/arXiv.2506.23714

Lan, L., Jiang, L., Yu, T., Liu, X., & He, Z. (2025). FullTransNet: Full Transformer with Local-Global Attention for Video Summarization. *Computer Vision and Pattern Recognition*. https://doi.org/10.48550/arXiv.2501.00882

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., & Zettlemoyer, L. (2019). BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. *arXiv*. https://doi.org/10.48550/arXiv.1910.13461

Li, H., Zhu, Y., Shang, Z., Wang, Z., & Wu, X. (2026). A comprehensive survey on video summarization: Challenges and advances. *IEEE Transactions on Circuits and Systems for Video Technology, 36*(1), 1216–1233. https://doi.org/10.1109/TCSVT.2025.3596006

Li, J., Li, D., Xiong, C., & Hoi, S. (2022). BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. https://doi.org/10.48550/arXiv.2201.12086

Li, W., Han, W., Man, H., Zuo, W., Fan, X., & Tian, Y. (2025). Language-Guided Graph Representation Learning for Video Summarization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. https://doi.org/10.48550/arXiv.2511.10953

Marevac, E., Kadušić, E., Živić, N., Buzađija, N., Tabak, E., & Velić, S. (2025). Multimodal Video Summarization Using Machine Learning: A Comprehensive Benchmark of Feature Selection and Classifier Performance. *Algorithms, 18*(572). https://doi.org/10.3390/a18090572

Narwal, P., Duhan, N., & Bhatia, K. K. (2022). A Comprehensive Survey and Mathematical Insights Towards Video Summarization. *Journal of Visual Communication and Image Representation, 89*(103670). https://doi.org/10.1016/j.jvcir.2022.103670

Nwakeze, O. M., & Mohammed, N. U. (2025). Design of a Secure Biometric Network Traffic Packet Data Transmission Using Deep Learning and Encryption Techniques. *International Journal of Computer Science and Mobile Computing, 14*(7), 108–123. https://doi.org/10.47760/ijcsmc.2025.v14i07.011

Okechukwu, O. P., & Okechukwu, G. N. (2026). A Multimodal Deep Learning Model for Emotion Recognition from Video, Audio and Text Data. *19*(2). https://doi.org/10.30574/wjaets.2026.19.2.0257

Pavan, G., Supriya, G., Rohith, K. S., Sriya, K. S., & Mohith, G. (2025). Automated Video Summarization Using BART, Whisper, and SBERT for Efficient Content Consumption. *International Journal for Modern Trends in Science and Technology, 11*(2), 65–73. https://doi.org/10.46501/ijmtst.v11.i02

Peronikolis, M., & Panagiotakis, C. (2024). Personalized Video Summarization: A Comprehensive Survey of Methods and Datasets. *Applied Science, 14*(4400). https://doi.org/10.3390/app14114400

Zhang, H. (2025). Bridging Multimodal and Video Summarization: A Unified Survey. *Proceedings of the 5th New Frontiers in Summarization Workshop* (pp. 157-171). Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.newsum-main.11

Zhu, S., Sun, L., Ma, Z., Li, C., & He, D. (2025). Prompt-supervised dynamic attention graph convolutional network for skeleton-based action recognition. Neurocomputing. *Neurocomputing, 611*(128623). https://doi.org/10.1016/j.neucom.2024.128623

Zhu, Y., Zhao, W., Hua, R., & Wu, X. (2023). Topic-aware video summarization using multimodal transformer. *Pattern Recognition, 140*(109578). https://doi.org/10.1016/j.patcog.2023.109578