MEEVS: Multimodal Educational & Explainable Video Summarizer

K Nithin, Dorai Sai Charan, Vijay Kumar, Priyanka Vivek
Department of Computer Science and Engineering
Amrita School of Computing, Bengaluru
Amrita Vishwa Vidyapeetham, India
kandinithin154@gmail.com, doraisaicharan09@gmail.com,
vijaygosu99@gmail.com, v_priyanka@blr.amrita.edu

***Abstract*—Multimodal Educational & Explainable Video Summarizer (MEEVS) is an end-to-end summarization system, which is developed to provide human-readable summaries from educational videos by combining both audio and visual textual content. In an effort to counterbalance the increasing demand for accessible and effective learning materials, MEEVS applies sophisticated natural language processing techniques to automate the summarization process. The system translates speech via the Whisper speech recognition model and captures on-screen text via Optical Character Recognition (OCR) via PaddleOCR with OpenCV-based preprocessing. These multimodal inputs are combined and passed through a transformer-based summarization model (BART), producing coherent and concise summaries appropriate for educational review. Besides text output, MEEVS provides downloadable summaries as PDFs and audio playback via Google's Text-to-Speech engine. User-accessible by design, Streamlit-based is used to guarantee usability by non-technical users. The three iterative sprints development, done in accordance with agile principles, was accompanied by extensive testing in each iteration. MEEVS is a full multimodal summarization tool and a real-time one that promotes learning efficiency by facilitating quick and accessible comprehension of lengthy educational materials.**

***Index Terms*—Multimodal summarization, Whisper speech recognition, Optical Character Recognition (OCR), BART transformer, educational video processing**

I. Introduction

Over the last decade, the explosive growth in online learning has made available an overwhelming array of video-based learning content on platforms like YouTube, Coursera, and Khan Academy. Although these resources provide unprecedented access to information, they tend to pose challenges regarding time and effort, accessibility, and content recall, especially for learners requiring rapid review or substitute modalities of understanding. Conventional summarization methods usually emphasize either written or oral content alone, not picking up the entire learning environment that is offered by both the verbal description and the on-screen written material like slides or notes.

Recent breakthroughs in artificial intelligence, especially for speech recognition, optical character recognition (OCR), and natural language processing (NLP), have made multimodal content understanding possible on a large scale. Although automatic transcription or slide extraction tools are available, few systems try to intelligently combine these sources to produce summaries that are both complete and understandable. Further, current summarization systems lack real-time generation capabilities and do not emphasize user accessibility through features such as audio summaries or a natural interface.

This work fills these gaps by presenting MEEVS (Multimodal Educational & Explainable Video Summarizer), an end-to-end multimodal summarizer that uses audio transcription, visual text extraction, and transformer-based NLP to produce coherent summaries from educational videos. The driving force of this project is to improve learning efficiency by minimizing the time and effort needed to comprehend lengthy educational materials, particularly for students, teachers, and visually/audibly challenged users. MEEVS seeks to make a contribution to the area by showing a practical and scalable method for real-time, explainable, and accessible multimodal summarization, thus pushing the boundaries of educational video content usability.

II. Literature Survey

There have been a large number of studies into a wide variety of text summarization models and methods, starting with conventional extractive methods like TextRank, TF-IDF, and fuzzy logic ranking mechanisms. These models concentrated on sentence rating based on statistical and linguistic characteristics to pull out salient content from a document [3], [7]. Initial neural models brought forth LSTMs and GRUs, with some of them incorporating semantic word embeddings (e.g., Word2Vec, GloVe) to gain a better contextual understanding. Seq2Seq models were the building blocks for abstractive summarization [4], whereas hybrid models such as EXABSUM brought in double pipelines for extractive and abstractive summarization based on word graphs and keyphrase-based reranking [2]. With the rise of transformer architectures, BART, T5, and GPT became strong models because of their ability to generate language, improving both

extractive and abstractive performance on benchmark datasets [6], [8].

Recent progress has extended to multimodal summarization, incorporating multiple data sources like text, images, audio, and video to generate more informative, context-sensitive summaries. The VATMAN model presents a Trimodal Hierarchical Multi-head Attention (THMA) mechanism to combine video, audio, and text through a Transformer-based model (BART), achieving performance improvement over unimodal and bimodal baselines [10]. Analogously, transformer-based method by Altundogan et al. combines automatic speech recognition with timestamped abstracts and pre-trained T5 and BART models fine-tuned for multimodal data, supporting extractive and abstractive summarization of audio and text inputs [11]. They use multimodal datasets and obtain higher ROUGE scores by incorporating audio semantics and temporal information. Furthermore, models such as Vision-SOGM and other multimodal approaches exhibit the significance of co-attention fusion and pretraining encoders for low-resource or domain-specific language domains [12], [9], [5].

In summary, the research has come a long way from shallow rule-based extractive summarizers to sophisticated transformer-based abstractive and multimodal designs. Notwithstanding progress, some gaps still exist: most models are still underperforming on long-form texts and non-robust in low-resource or cross-lingual environments [1]. Although transformer-based models such as BART and T5 lead research today [6], [8], very few methods take full advantage of multimodal synergy. Practical deployment is further hampered by interpretability, domain adaptation, and computational overhead issues. Lightweight multimodal models, deeper modality alignment, and domain-specific fine-tuning techniques are areas that future work needs to focus on for improving the real-world applicability and transferability of summarization systems.

III. Methodology

The growing popularity of accessible, compact educational material has created a need for smart systems to summarize long instructional videos. While a variety of tools are available for transcribing audio or extracting text from video slides, there are few systems that integrate both these modalities into a single summarization pipeline. To bridge this shortfall, we introduce MEEVS (Multimodal Educational & Explainable Video Summarizer)—an end-to-end, real-time system that utilizes both audio and visual text to create human-understandable and coherent summaries of learning videos. MEEVS marries Whisper, a state-of-the-art speech recognition model, with PaddleOCR for on-screen text extraction and combines them using a hybrid summarization approach that employs extractive summarization with BERT and abstractive summarization with BART models. The system is accessible, providing both text and audio output, and is deployed through an easy-to-use Streamlit interface. The next section describes the step-by-step methodology adopted in developing and deploying this system as shown in Fig1.

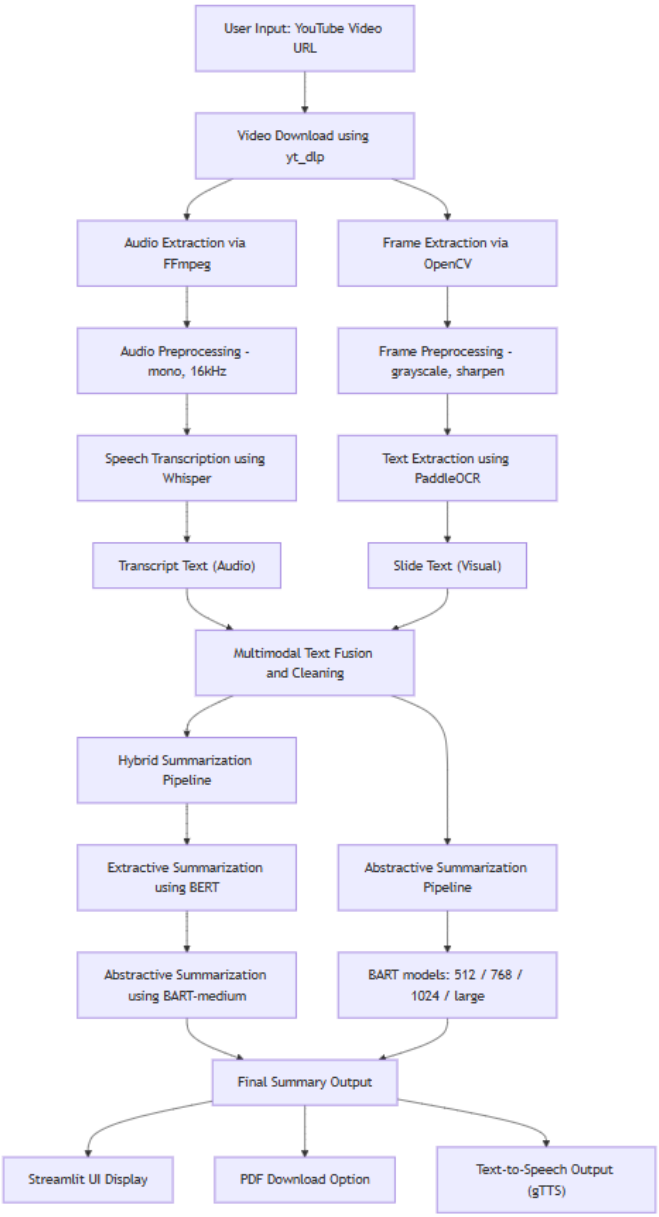

Fig. 1. System architecture diagram of MEEVS

A. Video Ingestion and Audio Processing

The system takes a YouTube URL as input via an interactive Streamlit interface. The video is downloaded in MP4 using yt_dlp with a filename generated from an MD5 hash of the URL. The downloaded video is subjected to audio extraction with FFmpeg, resulting in a mono-channel 16 kHz WAV file. This preprocessing allows for compatibility with the Whisper automatic speech recognition (ASR) model, which subsequently transcribes the audio content into plain text. This plain text transcript audio T_audio captures the audio narration of the video.

B. Frame Extraction and Visual Text Recognition

Concurrently, video frames are sampled every five seconds (up to 30 frames) with OpenCV. Frames are preprocessed by converting to grayscale, adjusting contrast, and performing binary thresholding. Text is extracted with angle classification on using PaddleOCR. The text output is filtered for minimum length and corrected with TextBlob to correct basic grammatical and spelling mistakes. This yields the visual text modality visual T_visual Typically based on slides, subtitles, or notes in the video.

C. Multimodal Text Fusion

The text outputs of the audio and visual processing phases are concatenated and cleaned with regular expressions. At this cleaning phase, symbols, citations and unnecessary characters are eliminated. The resulting multimodal document combined T_combined serves as input to the summarization module:

$$T_{\text{combined}} = \text{Clean}(T_{\text{audio}} + T_{\text{visual}}) \tag{1}$$

D. Extractive-Abstractive Hybrid Summarization Approach

To accommodate content coverage and abstraction, we have a hybrid summarization approach. We first conduct extractive summarization through a BERT-based model that picks the most semantically relevant sentences from the merged input. This is motivated by traditional extractive methods where sentence embeddings are ordered according to their contextual value. The resulting, extract T_extract, is then fed as input to one of a range of abstractive models.

E. Abstractive Summarization with BART Variants

We trained several variants of BART (Bidirectional and Auto-Regressive Transformers) on the ARXIV dataset from Hugging Face. The dataset comprises scientific abstracts and texts, which makes it optimal for summarizing educational material. The variants trained are: BART-medium models having maximum input lengths of 512, 768, and 1024 tokens, BART-large model with 1024 tokens. At summarization, the extractive output extract T_extract is tokenized and divided into overlapping chunks if necessary. Each chunk is summarized separately:

$$S_i = \text{BART}(T_i), \quad T_i \in \text{Chunks}(T_{\text{extract}}) \tag{2}$$

These intermediate summaries are concatenated and fed into the model once more to generate the final summary:

$$S_{\text{final}} = \text{BART}\left(\sum S_i\right) \tag{3}$$

This two-stage summarization (hierarchical abstraction) achieves coherence, completeness, and interpretability, with a reduction in redundancy. The employment of different token-length models also enabled us to contrast trade-offs in content coverage and output brevity.

F. Output Generation and Accessibility

The produced summary S_final is shown in the interface to view. The system provides extra accessibility features: Download as PDF with fpdf library, Play audio with Google Text-to-Speech (gTTS) to read the summary out instead. The whole system is designed with ease of use by non-technical users, for simple navigation and low setup.

G. Model Training and Evaluation

All BART models were fine-tuned on the arXiv scientific paper summarization dataset, training on different token lengths (512, 768, 1024) to compare model efficiency vs. output quality. BERT extractive was trained to optimize relevance of the content prior to abstraction. Standard summarization metrics like ROUGE, BLEU, and Content F1 were used for evaluation, with potential future work including BERTScore and human qualitative assessment.

IV. Results

From the Fig 2 the BART Large model consistently achieves higher ROUGE-1, ROUGE-2, and ROUGE-Lsum scores compared to Model 1, indicating better overlap with reference summaries in terms of unigrams, bigrams, and longest common subsequence.

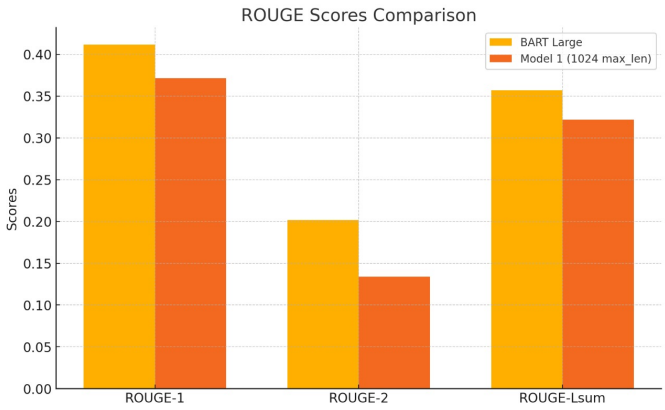

Fig. 2. Rogue score comparisons of BART large model and BART medium model with 1024 tokens

From the Fig 3 BART Large outperforms Model 1 on METEOR, showing it produces summaries that are more semantically and lexically aligned with the reference.

From the Fig 4 BART Large has higher precision, recall, and F1 scores in BERTScore, reflecting better semantic similarity and overall quality of generated summaries compared to Model 1.

This graph in Fig 5 displays all evaluated metrics for BART Large, highlighting its strong and balanced performance across ROUGE, METEOR, and BERTScore.

This graph in Fig 6 presents Model 1's metrics, showing comparatively lower scores across all evaluation metrics, suggesting it's less effective than BART Large in generating quality summaries.

TABLE I
OCR MODELS

S.No	Models	CER	WER	Exact Match Accuracy	Insertions	Deletions	Substitutions
1	Tr OCR	0.0210	0.0560	80.23%	3	6	8
2	Paddle OCR	0.0184	0.0426	93.23%	2	1	3

TABLE II
MODEL PERFORMANCE COMPARISON

S.No	Models	Rouge-1	Rouge-2	Rouge-L	Rouge-L Sum	Meteor	Bert-Precision	Bert-Recall	Bert-F1
1	BART Medium-512	0.3371	0.1143	0.2053	0.2951	0.1987	0.8648	0.8365	0.8502
2	BART Medium-768	0.3681	0.1297	0.2181	0.3171	0.2362	0.8623	0.8416	0.8516
3	BART Medium-1024	0.3715	0.1337	0.2188	0.3220	0.2388	0.8627	0.8423	0.8521
4	BART Medium-1024 Hybrid	0.3285	0.0895	0.1907	0.2796	0.1946	0.8456	0.8276	0.8363
5	BART Large-1024	0.4116	0.2019	0.3021	0.3571	0.2645	0.8855	0.8643	0.8746

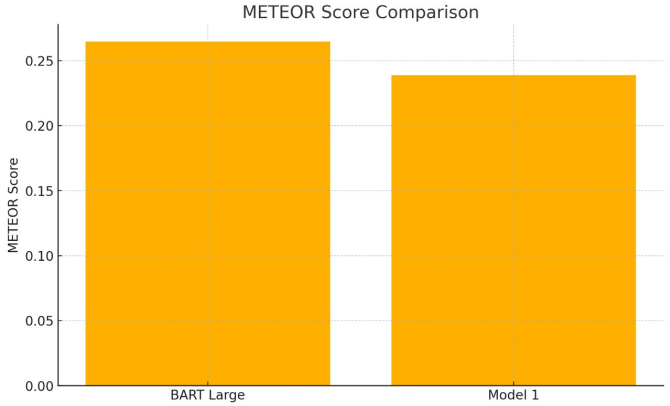

Fig. 3. Meteor score comparisons of BART large model and BART medium model with 1024 tokens

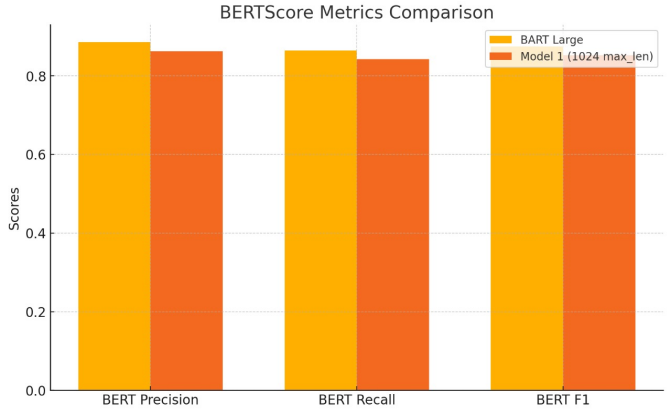

Fig. 4. BERT score comparisons of BART large model and BART medium model with 1024 tokens

The table I compares the performance of two OCR models—TrOCR and PaddleOCR—on Character Error Rate (CER), Word Error Rate (WER), and Exact Match Accuracy. PaddleOCR performs better than TrOCR on all measures, with a lower CER (0.0184 versus 0.0210) and WER (0.0426 versus 0.0560), in addition to having a much higher Exact

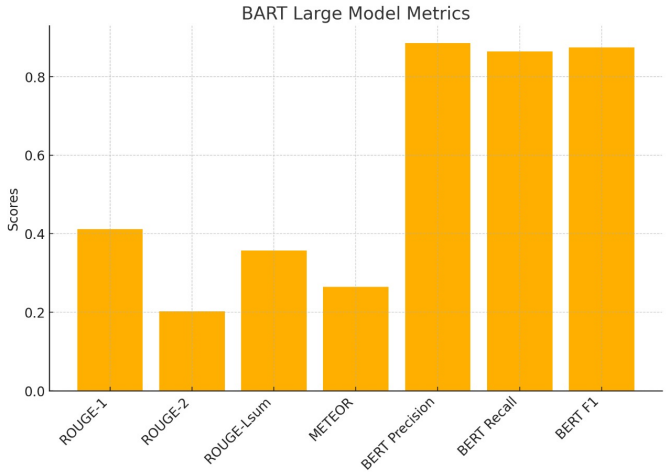

Fig. 5. Evaluation metrics comparision for BART large model

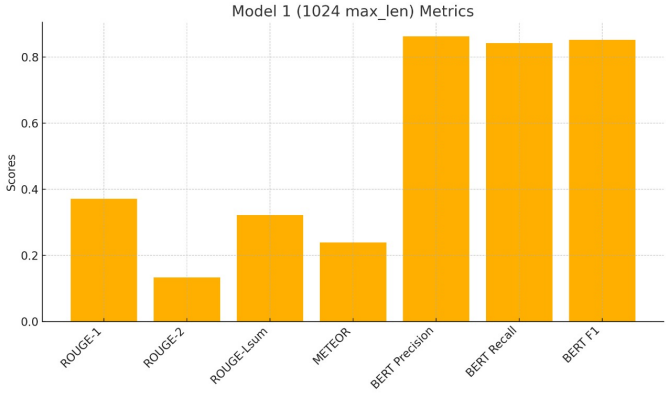

Fig. 6. Evaluation metrics comparision for BART medium model with 1024 tokens

Match Accuracy of 93.23% compared to 80.23%. In addition, PaddleOCR also results in fewer total errors, with fewer insertions, deletions, and substitutions. These results verify PaddleOCR as the more stable and accurate option for text extraction from video frames in this project.

The table II presents a comprehensive performance evaluation of various summarization models using ROUGE, METEOR, and BERTScore metrics. The BART-Large-1024 model achieves the best results overall, with the highest ROUGE-1 (0.4116), ROUGE-2 (0.2019), and BERT F1 Score (0.8746), indicating superior content coverage and semantic accuracy. Among the medium variants, the BART-Medium-1024 model performs better than the 768 and 512 versions. Interestingly, the hybrid model (BERT + BART-Medium-1024) provides competitive results, demonstrating the effectiveness of combining extractive and abstractive techniques for long-form educational content.

V. Conclusion

The summarization pipeline was implemented in both abstractive (BART variants) and hybrid extractive-abstractive (BERT + BART) approaches. For the tested models, BART-Large-1024 provided the best performance with a ROUGE-1 score of 0.4116 and BERT F1 score of 0.8746, which confirms its superior capacity for preserving and abstracting lengthy educational content. However, the hybrid model is relatively in the relationship to abstraction but remains quite competitive while incurring lower computational complexity, which highlights its utility for lighter, real-time applications. One of the most important strengths of this project is its modular architecture, real-time modality, and accessibility features such as downloadable PDF summaries and text-to-speech output. These make MEEVS extremely usable and scalable as a tool for learners, educators, and content consumers. Some limitations were also revealed, such as longer inference time for large models, absence of end-to-end multimodal attention integration, and possible sensitivity to noisy audio or complicated slide visuals. Future work may investigate adding unified multimodal transformer architectures (e.g., with cross-modal attention) to enhance real-time performance for low-resource settings, as well as adding support for multilingual content. Human grading and task-specific tuning can further enhance the readability and relevance of summaries. MEEVS thus presents a useful and applicable contribution to the domain of educational content summarization, closing the loop between multimodal data processing and human-focused learning experiences.

References

[1] Supriyono, & Wibawa, Aji & Suyono, & Kurniawan, Fachrul. (2024). A survey of text summarization: Techniques, evaluation and challenges. Natural Language Processing Journal. 7. 100070. 10.1016/j.nlp.2024.100070.

[2] Alami Merrouni, Z., Frikh, B. & Ouhbi, B. EXABSUM: a new text summarization approach for generating extractive and abstractive summaries. J Big Data 10, 163 (2023). https://doi.org/10.1186/s40537-023-00836-y

[3] Dhanda, Namrata & Gupta, Kapil. (2024). A Novel Approach to Text Summarization Using Machine Learning. Asian Journal of Research in Computer Science. 17. 95-104. 10.9734/ajrcos/2024/v17i4432.

[4] Gangundi, R., Sridhar, R. IWM-LSTM encoder for abstractive text summarization. Multimed Tools Appl 84, 5883–5904 (2025). https://doi.org/10.1007/s11042-024-19091-1

[5] Khilji, A.F.U.R., Sinha, U., Singh, P. et al. Multimodal text summarization with evaluation approaches. Sādhanā 48, 226 (2023). https://doi.org/10.1007/s12046-023-02284-z

[6] Rao, R., Sharma, S. & Malik, N. Automatic text summarization using transformer-based language models. Int J Syst Assur Eng Manag 15, 2599–2605 (2024). https://doi.org/10.1007/s13198-024-02280-4

[7] Deo, Satya, Debajyoty Banik, and Prasant Kumar Pattnaik. "Customized long short-term memory architecture for multi-document summarization with improved text feature set." Data & Knowledge Engineering 159 (2025): 102440.

[8] Onan, Aytuğ & Alhumyani, Hesham. (2024). DeepExtract: Semantic-driven extractive text summarization framework using LLMs and hierarchical positional encoding. Journal of King Saud University - Computer and Information Sciences. 36. 102178. 10.1016/j.jksuci.2024.102178.

[9] S. Rafi and R. Das, "Abstractive Text Summarization Using Multimodal Information," 2023 10th International Conference on Soft Computing & Machine Intelligence (ISCMI), Mexico City, Mexico, 2023, pp. 141-145, doi: 10.1109/ISCMI59957.2023.10458505.

[10] D. Baek, J. Kim and H. Lee, "VATMAN : Video-Audio-Text Multimodal Abstractive Summarization with Trimodal Hierarchical Multi-head Attention," 2023 14th International Conference on Information and Communication Technology Convergence (ICTC), Jeju Island, Korea, Republic of, 2023, pp. 1475-1478, doi: 10.1109/ICTC58733.2023.10392391.

[11] T. G. Altundogan, M. Karakose and S. Tanberk, "Transformer Based Multimodal Summarization and Highlight Abstraction Approach for Texts and Speech Audios," 2024 28th International Conference on Information Technology (IT), Zabljak, Montenegro, 2024, pp. 1-4, doi: 10.1109/IT61232.2024.10475775.

[12] Y. Song, N. Lin, L. Li and S. Jiang, "A Vision Enhanced Framework for Indonesian Multimodal Abstractive Text-Image Summarization," 2024 27th International Conference on Computer Supported Cooperative Work in Design (CSCWD), Tianjin, China, 2024, pp. 61-66, doi: 10.1109/CSCWD61410.2024.10580245.