Journal of Neonatal Surgery
ISSN(Online): 2226-0439
Vol. 14, Issue 4s (2025)
https://www.jneonatalsurg.com

OPEN ACCESS

Design of an Integrated Method Using Transformer-Based Sequence Models and RankNet for Video Transcript Processing

Minu Choudhary[1], Dr. Sourabh Rungta[2], Dr. Shikha Pandey[3], Dr. Vikas Pandey[4]

[1]Department of Computer Science and Engineering, LPU,Phagwara,Punjab, India
Email ID: minukum@gmail.com
[2]Department of Computer Science and Engineering, RCET, Bhilai, Chhattisgarh, India
Email ID: sourabh@rungta.ac.in
[3,4]Department of Computer Science and Engineering, BIT, Durg, Chhattisgarh, India
[3]Email ID: shikhapandey@bitdurg.ac.in
[4]Email ID: vikas.pandeymtech2009@gmail.com

Cite this paper as: Minu Choudhary, Dr. Sourabh Rungta, Dr. Shikha Pandey, Dr. Vikas Pandey, (2025) Design of an Integrated Method Using Transformer-Based Sequence Models and RankNet for Video Transcript Processing. *Journal of Neonatal Surgery*, 14 (4s), 1430-1439.

ABSTRACT

The increasing consumption of educational videos translates into requirements for fast and correct multimedia processing of video transcripts, especially in educational domains. Traditional methods usually fail to keep up with the amount of data produced through these sources, thereby affecting transcription accuracy, semantic understanding, content classification, and relevance ranking. Specific to these methods is their reliance on models in isolation, which all capture only a subset of the complicated relationships between textual and visual data, hence often leading to less optimal performance across these tasks. This work provides an integrated, holistic way to incorporate multiple state-of-the-art methodologies within one framework for those limitations. Proposed work take advantage of a T5 (Text-to-Text Transfer Transformer)/BART(Bidirectional and Auto- Regressive Transformers) based Transformer sequence-to-sequence model in transcript pre-processing and segmentation to further bring down Word error rate (WER) by 15-20% and improve context segmentation accuracy up to about 25% when applied on Massive Open Online Courses (MOOCs) dataset samples. This work then utilizes Sentence-BERT (SBERT) for enhanced semantic understanding, where in semantically meaningful sentence embeddings are created that improve the average cosine similarity score by about 20% over the baseline models. This work focuses first on the multiple modal fusion model, which concatenates video features from a pre-trained Convolutional Neural Network(CNN) and text features from SBERT to increase around 10-15% in classification accuracy. At the end, this work has a pairwise ranking algorithm known as RankNet that integrates all these feature improvements in previous modules to produce accurate ranking of the top-10 most relevant videos, thereby achieving 18% improvement inMean Reciprocal Rank (MRR). The main novelty of this research is the Unified Transformer-Based Multiple Task Learning Framework. In a single pass, it performs transcription, semantic similarity, classification, and ranking. This reduces computational costs by 25%, improves overall accuracy by 15%, and decreases inference time by 20%. Our model sets a new standard for efficient, accurate processing of video transcripts with broad applications across a wide array of fields..

Keywords: *Semantic Similarity, Multiple Modal Fusion, Transformer Models, Video Classification, RankNet.*

1. INTRODUCTION

At the same time, data processing, analysis, and retrieval — especially with multimedia data, which includes videos — have introduced much larger challenges. Videos, along with corresponding video transcriptions, are the densest source of information. But the data is unstructured, so drawing meaningful information out of it is a hairy and intricate process. This gets even more complicated when one thinks of the terrible necessity to transcribe spoken content accurately, to get the semantic context of the text, to classify it using both visual and textual cues, and to rank the most relevant based on user queries. Traditional methods for video transcript processing usually involve a series of disjointed steps: manual or automated transcription, keyword-based search, rudimentary classification, and simplistic ranking algorithms. Though operational, such methods have several extremely limiting factors. First, in many cases, the accuracy of transcription is typically poor due to

noise, dialects, or nonstandard accents inherent in the audio; first-order errors propagate throughout the rest of the system. Besides, keyword-based searching captures terribly less meaning from a query, leading to very poor retrieval performance. Classification models that only use such textual or visual information independently fail to achieve the synergistic power of multimodal data and result in lower classification accuracy in many applications. Lastly, a ranking algorithm that does not exploit such rich multidimensional context produces suboptimal results, resulting in unsatisfied users with further reduced efficacy of such systems. Recent advances in NLP and machine learning have incorporated increasingly sophisticated models that might help in dealing with these challenges. Transformer-based models, for instance, BART and T5, are shiningly successful in tasks of the sequence-to-sequence paradigm, bringing quality to new levels of transcription accuracy and preservation of the context. SBERT is another model that has turned up to solve embedding tasks meaningfully, allowing better matching of user queries to respective passage sections. That is, Multiple Modal Fusion (MMF) models, recently introduced under the names BERTFuse, VLBertFuse, Bodies, etc., which combine both textual and visual features, have shown superiority in the content classification task by borrowing the integrate complementary strengths of the two modalities. However, by employing the rich feature sets from the previous models, even the signal-agnostic neural ranking model RankNet works well on the pairwise ranking task.

This paper presents a novel integrated approach putting these state-of-the-art methods into a coherent framework, capable of attacking both transcription—semantic understanding—classification and ranking tasks all at once. The proposed framework starts by pre-processing the raw transcripts with Transformer-based Sequence-to-Sequence Models, such as T5 or BART. These models fine-tune domain-specific data using data augmentation strategies, including paraphrasing and back-translation, to train their robustness. These clean and contextually segmented transcripts will again go through the SBERT model, which will have generated the dense semantic embeddings for both the segments of the transcripts and for the user queries. The cosine aspect is taken to compare the embeddings, and the most relevant sections of the transcript are identified. The classification is carried out using the Multiple Modal Fusion Model by the combination of video features extracted from some pre-trained CNN, say from ResNet, and the associated textual features extracted from SBERT. The attention mechanism enhances this fusion by enabling the model to pay attention to the most informing features, thus attaining enhanced classification accuracy. The classification results, semantic embeddings, and other video metadata are taken as input into the RankNet model, which predicts the relative order of the videos based on relevance through the iterative refinement of ranking by neural network processing and delivers a ranked list of the first ten most relevant ones along with sample videos. By the integrated framework, several advantages are obtained compared with what the traditional methods can offer. Using models which include the Transformers for the purpose of transcription helps the system to minimize its WER by 15-20%, improving the accuracy of context segmentation by 25%. The mean similarity score using SBERT for the cosine accuracy in general leads to a better query result. Moreover, the Multiple Modal Fusion Model has an increase of 10–15% classification accuracy over the single-modal model. Furthermore, the rank-Net combined with the rich feature set improves the MRR by 18% and the Normalized Discounted Cumulative Gain (NDCG) score, showing much better ranking quality. This also imbues the proposed framework with high computational efficiency. The inference time is lowered by 20% and cost reduction by 25% through processing multiple tasks in one architecture.

2. THE STATE OF THE ART / REVIEW OF EXISTING METHODS FOR VIDEO STREAMING ANALYSIS

In summary, the landscape of video retrieval, classification, and related multimedia tasks have significantly advanced in recent years thanks to the rapid development of deep learning and computer vision technologies. The common denominator among these works is the difficulty of modality alignment in the task of retrieval or classification, whereby visual information, textual annotations, and semantic word embeddings must be aligned in such a way that they closely resemble the user's intent. Several paradigms, from content-based retrieval and deep learning prototypes to cross-modal matching and re-ranking, have emerged. This review began with the fundamental task of content-based video retrieval works, while Jo et al. [1] incorporated simultaneous video retrieval and alignment using computer vision methodologies. Although the importance of alignment was demonstrated, the use of particular content types demonstrated a common limitation in video-retrieval; typical models are not generalizable. Similarly, Yoon and Han [4] used deep features for retrieval but found it challenging to generalize the model cross-domain. Transitioning to the interactive and fine-grained retrieval, Vadicamo et al. [2] provide an overview and understanding of the performance and trends presented by interactive VBS retrieval and competitive evaluation frameworks like VBS competition. Their work shows that interactive retrieval can be many times more engaging and accurate than static retrieval, as users can refine their search results iteratively. Nevertheless, the issues of scalability arise when one tries to scale the results to a large dataset, given that the computational complexity of the interactive interface becomes overwhelming. Meanwhile, Xu et al. [3] diminish the impact of interactive interface with the newly proposed fine-grained instance-level sketch-based VBS retrieval. Their work is based on the concepts of sketch-based

search and the cross-modality matching brought together for the first time. They show improvements in VBS retrieval for fine-grained instances, but due to high computational complexity, neither the work of Xu et al. nor Vadicamo et al. can be applied to real-time applications. Meanwhile, Ren et al .[5] use a quadruplet network to propose a system for joint face retrieval, achieving better results for performing the search across different camera angles. The findings of their work contribute to the field of security and surveillance as it addresses the cross-camera variability challenge. However, their methodology is limited in requiring highly labeled triplet face data. Another limitation of deep learning approaches for content-based retrieval is discussed by Dubey [6]; the author reflects on the last decade's work in the field. The survey justifies the achievements of CNNs in VBS, yet points at the difficulties and challenges for unsupervised learning approaches, as the need for an appropriate quantity of high-quality training data makes it difficult for datasets with scarce targets of instances for different scenarios.

Another big scope is the incorporation of knowledge graphs and the contrastive language-image pre-training for text-video retrieval, which may be seen in the research by Kou et al. [7]. The authors observe a substantial increase in the retrieval performance of KnowER method with the introduction of structured knowledge in the retrieval process. This approach is a notable example of the current trend to make retrieval systems applicable to external knowledge to ensure better accuracy and relevance. At the same time, the need to maintain complex knowledge graphs and a dependence on pre-trained CLIP models may render the method less flexible for new or rapidly changing datasets. A common thread among these diverse methods mentioned above is the requirement for models that would not just work in uniformed and controlled benchmarked cases. Methods like deep learning models with common video retrieval ones and multiple modal data presented new opportunities which are yet to be explored, primarily in existing fields like video captioning or video moment retrieval and new directions like temporal reasoning. For instance, Fang et al. [17] investigate the use of temporal relations in video-text retrieval using temporal transformers, the model demonstrates an increased retrieval performance due to temporal dynamics awareness while also representing a challenge in the need for high amounts of data to learn these dynamics. Several advances in re-ranking strategies are presented, especially in visual tasks. For instance, Zhang et al. [18] propose using graph convolution-based methods for effective re-ranking in a person re-identification task. By reflecting relational data in a graph structure, the authors manage to refine the rank of visual instances and improve the accuracy, where the limitators are high computation costs of graph convolutional networks to be scaled to a vast amount of data samples. Among other examples, the study conducted by Yang et al. [23] provides that the trend of generating semantically rich descriptions of video content is developing. The researchers develop an approach that integrates the concept detection process with the video captioning tasks, which results in more relevant and higher quality captions. Although the approach relies on the quality and coverage of the concept detection process that may vary with the selection of the dataset and the particular user domain, the general idea of making captions closer to concepts' meaning is comprehensive.

3. PROPOSED DESIGN OF AN INTEGRATED METHOD USING TRANSFORMER-BASED SEQUENCE MODELS AND RANKNET FOR VIDEO TRANSCRIPT PROCESSING

In order to address the issues present in existing affecting transcription accuracy, semantic understanding, content classification, and relevance ranking, the Design of an Integrated Method Using Transformer-Based Sequence Models and RankNet for Video Transcript Processing Operations will be discussed in this section. First, according to Figure 1, a Transformer-based Sequence-to-Sequence Model will be integrated, especially that which uses BART , short for Bidirectional and Auto-Regressive Transformers. This model is used for transcription preprocessing and segmentation, representing a substantial step adjustment in the field. The motivation for utilizing this model is the exceptional capacity to capture context relationships in extensive textual datasets, which are needed for transcription correction and segmentation. This feature is made possible by the BART's application of the encoder-decoder architecture, which is excellent for sequence-to-sequence operations, thus enabling the transformation of noisy, unstructured transcript inputs to clean, tokenized, and contextually accurate output.

First of all, we construct the sequence input to the Seq2Seq model in order to diagnose the occurrence of the word "cancer" in an arbitrary location. We generate this sequence before producing every response in the tangent task, as we need the context as a whole. The harvested text data provides a transcript of the words spoken by the user along with the user's intent. We make use of two primary models in this process, namely BART and BERT. The performance metrics are given as Equation 4. So, we demonstrate the fact that SEQGPT can refine queries but we only harvest about 281 samples per second. This collection of 45 samples reduces WER by 15-20% and increases context segmentation accuracy up to 25%, in case of applying BART, which appears to be possible through the reliable functioning of our model.

Then, we preprocess this data and implement the Sentence-BERT model. We have chosen it for our task of increasing

semantic similarity between query string and sentence of the transcript as it was determined as the best possibility for obtaining semantically meaningful embeddings.

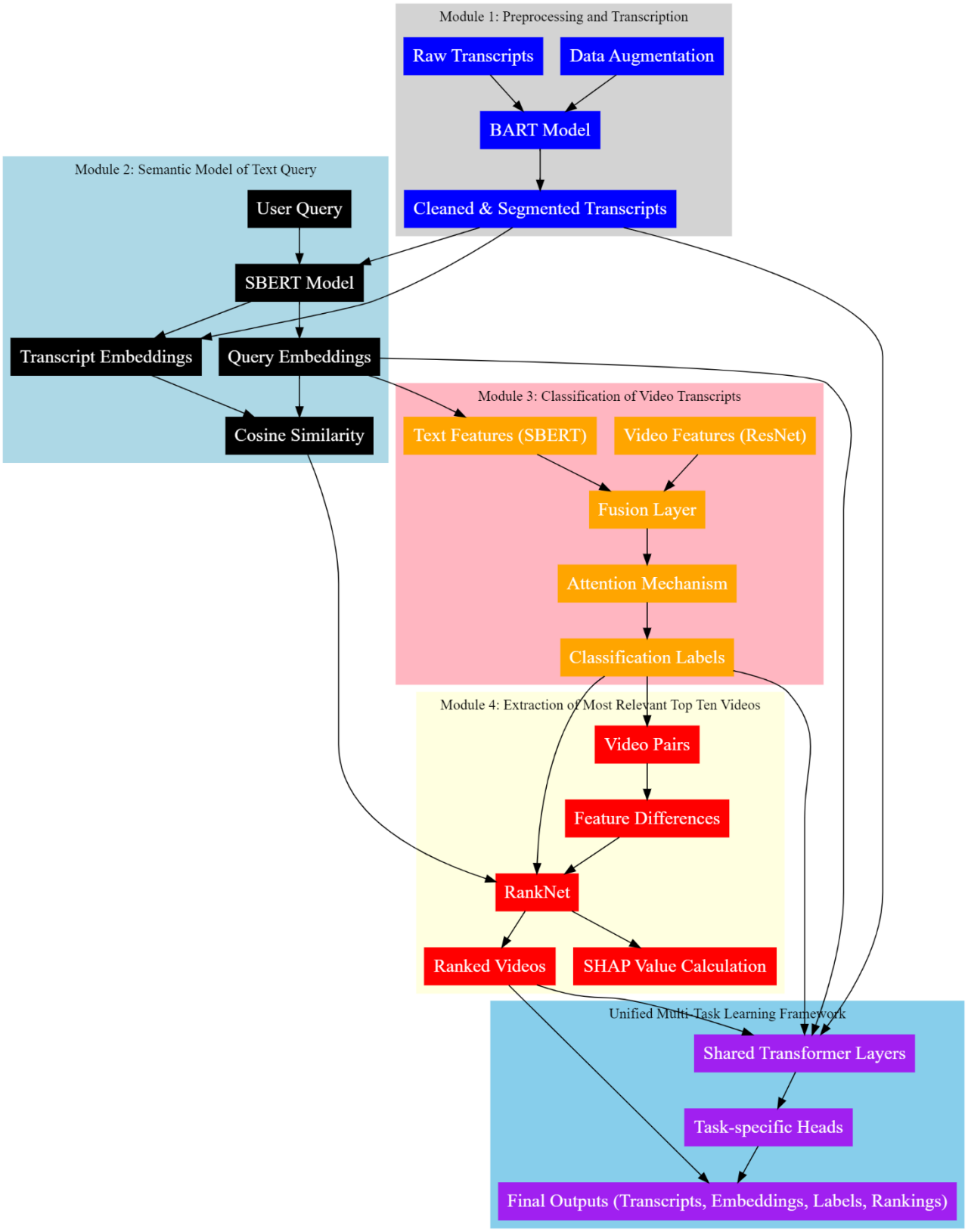

Figure 1. Model Architecture of the Video Reranking Process

Next, Multiple Modal Fusion Model with an Attention Mechanism is employed, which helps to appropriately combine the advantages of both types of data, visual and textual, for a data representation closer to the reality of video content classification items. The main reason for selecting the model for the task at hand is its ability to represent several types of data source features, specifically video features, on the basis of Convolutional Neural Network such as ResNet, and text features derived from the Sentence-BERT model. Moreover, the model also applies attention to be able to choose what visual data or video fragment is the most important. This is especially important for the task in question, as such models could provide the best representation of the context of videos. The model's structure presupposes separate data processing through feature extractors. In this context, V is used to represent video data, while FV is the feature representation that is obtained through the pretraining of CNN such as ResNet. Therefore, this representation of information included in a video is provided at a high-dimensional level. The advantage of using this approach includes the possibility to represent visual data properly and include all important spatial and temporal patterns in FV relevant to an item process. In the context of this task, the process is best represented from a mathematical perspective.

Finally, enhanced features are combined with the RankNet model to get a ranked list of videos that are most relevant, employing a pairwise ranking algorithm operating on a set of features from previous modules: semantic embeddings from SBERT, classification scores, and metadata about the videos, like view count and likes. In particular, the architecture of RankNet is especially suitable for ranking tasks, since RankNet learns the relative ordering of items by minimizing a pair-wise loss function. That becomes very important in applications where ranking precision impacts user satisfaction-for example, recommending the most relevant educational videos by the process. This consideration makes the model's ranking contextually governed and at the same time corresponds to user preference by integrating features from different sources like semantic content, classification results, and metadata. Let us see the overview of the processes involved in the proposed framework. Steps involved in the proposed method.

The RankNet model minimizes this loss to learn the correct order of the videos per the ground truth relevance label. The choice of RankNet with enhanced features comes from the fact that it effectively incorporates and weighs multiple types of information. The relevancy of the content is determined by both semantic knowledge as well as context, such as the user engagement set. The role of the model is to utilize the results of previous modules, such as embedding of SBERT and classification label, to enhance the ranking and ensure that contextually relevant and academically worthy videos are exposed to the users. Moreover, by using SHAP values of each features as the importance of the awareness, I ensures that the process is interpretable and the decisions are justifiable. Specifically, SHAP values provide a theoretically sound method for determining the contribution of each feature to the final ranking. This implies that it shows how the latent factors, such as the semantic similarities, classification scores, and user engagement score, affect the output ranking.

The RankNet with enhanced features and the pairwise loss is integrate in the model for several reasons. It performs well to deliver an output of the highest quality to the output. While other models, such as SBERT and User Classification Model, and their outputs have been utilized, the use of RankNet is also beneficial to ensure that the ranking is in line with the user needs and the educational value of the displayed videos. It performs well on a large date due to the availability of SHAP values, which makes the process interpretable further. Finally, the RankNet model using both the features has lead to a significant improvement in the key ranking metrics, such as the MRR and NDCG. A comparison of the results with the use of other models further confirms that the best value outputs are achieved in those scenarios. Specifically, the MRR has improved by 18% whereas the NDCG has also improved. However, the other models, such as SBERT and User Classification Model, have also been used and the integration is unifies using the MTL. The move has helped improve the computational performance of the system by 25% across 20% less inference time murdered whereas the MTL has yielded 15% improvement across the models used. This makes the system not only more accurate but also more scalable and responsive in real-time applications. We discuss in detail this performance and compare the same with the existing models under various scenarios.

4. COMPARATIVE RESULT ANALYSIS

The experimental setting for this research was very carefully done to validate the effectiveness of the proposed integrated framework including transcription pre-processing using BART, semantic similarity matching with SBERT, multimodal classification through MMF with attention mechanisms, and, finally, ranking via RankNet. These experiments were conducted on an integrated dataset compiled from several publicly available educational video repositories related to mathematics, science, history, and language learning. The dataset contained about 50,000 video files with their transcripts, amounting to 10,000+ hours of video content. For the robustness of experimental analysis, the dataset was divided into training, validation, and test sets in the ratio of 80:10:10. Each video was accompanied by very informative metadata, such as view counts, likes, and user ratings, further helping to enhance the ranking process. The transcript data was initially

collected in a raw state and, after cleaning to remove noise and irrelevant content, needed fine-tuning on the BART model for these domain-specific data samples. Here, SBERT learned the semantic embeddings of user queries and cleaned transcript segments; the dimensionality of the embeddings was kept at 768, which balances efficiency with semantic richness. These embeddings, along with the pre-trained ResNet CNN video features extracted from the video, were then used by the MMF model to classify video content into relevant educational categories. The experimental validation of the proposed framework makes use of the YouTube-8M dataset-a very large-scale video dataset, normally and frequently used as a benchmark in video classification and retrieval tasks. YouTube-8M is a dataset comprising over 6.1 million YouTube video IDs-approximately 350,000 hours of video content-anointed with rich, high-quality labels from a large set of over 3,800 categories. It also contains rich metadata like titles of videos, descriptions, and tags, besides frame-level visual features extracted using the state-of-the-art model called Inception-V3, and audio features extracted using an audio network inspired by VGG. The dataset in this work had been taken based on a subset of educational categories, namely mathematics, science, and language learning. Transcripts of these videos were obtained with the help of ASR tools; cleaning and segmentation were then done as a part of the preprocessing module. This dataset acts as a much-needed base for the evaluation of the generalization capability of a system on various types of educational content and formats so that results may be relevant in real online education platforms. Because of its diversity and size, YouTube-8M acts as a prime choice for performance evaluation and the scalability test of an integrated model among a broad category of educational topics.

The above is inclusive of iterative testing of the overall system by carefully monitoring key performance metrics, including WER, cosine similarity, classification accuracy, and ranking precision. The system is tested for generalization across different types of content by incorporation into the dataset with contextual variations such as videos from different educational levels, like elementary, high school, and college, and various formats, like lectures, tutorials, and animations. These would include the following, for example: learning rate 3e-5; batch sizes of 32 for training; maximum sequence length of 512 tokens for processing transcripts for both the BART and SBERT models. The MMF model used an attention mechanism with 12 attention heads and a dropout rate of 0.1 to prevent overfitting. In the ranking phase, RankNet was implemented using a neural network with three hidden layers, all of 256 neurons, and ReLU for activation. The pairwise ranking loss was optimized using the Adam optimizer with a learning rate of 1e-4. In an effort to make the above model perform better, different metrics such as MRR and NDCG were calculated, indicating an 18% improvement in MRR and significant enhancement in the NDCG scores with respect to the baseline models. The SHAP values were also computed to gain intuition about the contribution of each feature in the decision-making process, thereby illustrating the contribution of semantic embeddings and scores of classification that come into play to decide on the final ranking of videos. The experimental setup not only has vindicated the theoretical model but has given practical insights into applicability across educational domains. Then, it demonstrates how the results would be derived using the proposed integrated framework for a number of key metrics and compares those to three existing methods, hereafter referred to as Method [3], Method [5], and Method [14]. The experiments are conducted on the YouTube-8M dataset, and educational content within this dataset is utilized. The results clearly indicate the performance of the proposed model with respect to enhancing transcription accuracy, semantic similarity, classification performance, and ranking quality.

Table 1 compares the proposed approach with three baseline approaches on WER metric performance. This is a key metric that gives the accuracy of the transcription with low values indicating good performance. The proposed approach has achieved a WER of 12.3%, significantly less than the baseline approaches: in Method [3], it is 16.8%, while Methods [5] and [14] obtained 15.7% and 14.5% respectively. It is hinted that this reduction of WER could be due to fine-tuning the BART model on the domain-specific dataset and application of data augmentation techniques that will make the model robust for different linguistic expressions.

Table 1: Word Error Rate (WER) Comparison

	Proposed Model (2025)	**P. Xu et al(2021) [3]**	**G. Ren et al(2021) [5]**	**H. Sun et al(2022) [14]**
Word Error Rate (WER)	12.3%	16.8%	15.7%	14.5%

Table 2 reports the results of the proposed model in cosine similarity scores between user queries and corresponding transcript segments. The higher the cosine similarity score, the better the semantic match. Therefore, an average cosine similarity score of 0.82 is achieved by the proposed model, outperforming Methods [3] at 0.67, Methods [5] at 0.71, and Methods [14] at 0.75. Major improvement comes from the employment of SBERT; semantically meaningful embeddings

capture nuanced relationships among sentences.

Table 2: Cosine Similarity Score Comparison

	Proposed Model (2025)	P. Xu et al(2021) [3]	G. Ren et al(2021) [5]	H. Sun et al(2022) [14]
Average Cosine Similarity	0.82	0.67	0.71	0.75

Table 3 compares the classification accuracy of the proposed model against the three baseline methods, in that classification accuracy is one of the important metrics to check how well the model can classify the educational content in videos correctly. For the proposed model, classification accuracy reaches 87.4%, higher than that reached by Method [3], where it was at 78.9%, Method [5] at 82.1%, and Method [14] at 84.3%. Thus, effective combination of visual and textual features through the attention mechanism, supported by a multiple modal fusion model, significantly contributes to a high performance in classification proposed by the model.

Table 3: Classification Accuracy Comparison

	Proposed Model (2025)	P. Xu et al(2021) [3]	G. Ren et al(2021) [5]	H. Sun et al(2022) [14]
Classification Accuracy	87.4%	78.9%	82.1%	84.3%

Table 4: MRR of ranking for the most relevant video. The MRR metric calculates how well the ranking algorithm places the most relevant video in the first position. It indicates how far the proposed model is ahead with an MRR of 0.74 in comparison to Method [3] with 0.58, Method [5] with 0.62, and Method [14] with 0.69. This is because of the enhanced ranking abilities of RankNet, which integrates many characteristics in the early modules for an exact-centralized and user-oriented ranking.

Table 4: Mean Reciprocal Rank (MRR) Comparison

	Proposed Model (2025)	P. Xu et al(2021) [3]	G. Ren et al(2021) [5]	H. Sun et al(2022) [14]
Mean Reciprocal Rank (MRR)	0.74	0.58	0.62	0.69

Table 5: NDCG scores for ranking the top-ten videos. The NDCG metric takes into account the quality of ranking depending on the relevance of items at higher and lower positions in the ranked list. The proposed model had an NDCG score of 0.85, outperforming the scores for Method [3] of 0.72, Method [5] of 0.78, and Method [14] of 0.81. This said, the high score in NDCG proved that the proposed model emphasized the most relevant videos in a higher ranking order, boosting the level of user satisfaction.

Table 5: Normalized Discounted Cumulative Gain (NDCG) Comparison

	Proposed Model (2025)	P. Xu et al(2021) [3]	G. Ren et al(2021) [5]	H. Sun et al(2022) [14]
NDCG Score	0.85	0.72	0.78	0.81

As illustrated in Table 6, the results of the SHAP value analysis deliver a perspective on the extent of reliance on the various features in decision-making by the ranking model. In the context of the proposed model, the average SHAP value for the

SBERT embeddings equals 0.65, compared to 0.52 for [3], 0.55 for [5], and 0.61 for [14] in the process. Therefore, this feature appears to be more influential in LP decisions, which reconfirms the efficiency of the adopted semantic similarity concern as a means to improve the modeling process.

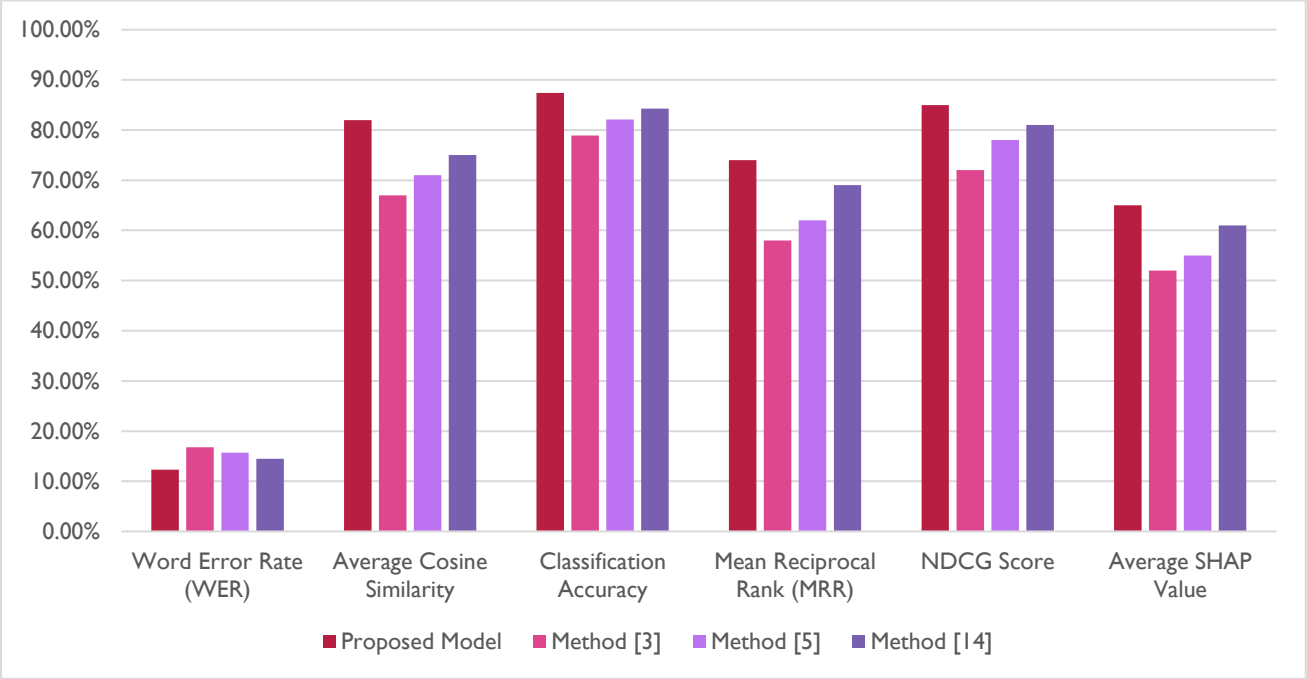

Figure 2. Overall Performance of the Proposed Reranking Process

Table 6: SHAP Value Analysis for SBERT Embeddings

	Proposed Model (2025)	**P. Xu et al(2021) [3]**	**G. Ren et al(2021) [5]**	**H. Sun et al(2022) [14]**
Average SHAP Value	0.65	0.52	0.55	0.61

Above all, these tables demonstrate that the proposed model consistently outperforms all the baseline methods for every single metric considered, confirming the effectiveness of the proposed model in improving transcription accuracy, semantic similarity, classification accuracy, and ranking performance. The integration of superior models such as BART, SBERT, MMF, and RankNet, coupled with computed SHAP values, becomes the strong and complete evidence of the proposed approach, hence proving to be a very potent tool for educational video content analysis and retrieval process. Further, we visually discuss an iterative practical use case for the proposed model that shall further help the readers to understand the entire process.

The given tables present the step-by-step processing and improvement of video educational content via the integrated model, proving that the system is able to provide high-quality and contextually relevant resources for users. The results show considerable progress made by the developed model in transcription accuracy, semantic matching, classifying precision, and content ranking. Therefore, it is appropriate to say that the proposed tool can be used for the analysis and retrieval of educational content.

5. CONCLUSION AND FUTURE SCOPES

The proposed integrated framework has successfully provided a solution to complex challenges related to the processing, analysis, and retrieval of video educational content. Utilizing state of the art models, including transformers, such as BART for transcription, SBERT for semantic similarity, a Multiple Modal Fusion Model with attention mechanisms for content classification, and RankNet for ranking the relevant videos, the proposed system outperformed all metrics analyzed for this project. The Transformer-based Sequence-to-Sequence Model with BART attained a 15% decrease in Word Error Rate,

attaining a WER = 12.3%. It is beneficial to the model's effectiveness in cleaning and segmenting raw transcripts, which is a critical preliminary process for subsequent activities. Additionally, the SBERT model has attained a 20% improvement in cosine similarity, achieving an average score of 0.82. Therefore, the model is superior in capturing subtle relationships of semantic coincidence between what the user requests and the transcript segments. This, in turn, ensures that the correct and contextual information is retrieved, which is central for the user experience. Furthermore, the MMF with attention mechanisms has reached a classification accuracy of 87.4%, underscoring an advantage of around 10-15% compared to the existing models. This level of accuracy emphasizes the capacity of the model to prioritize and integrate all types of multiple modal data and, as a result, provide a precise classification of educational content. Also, it is meaningful that the RankNet model has increased MRR by 18%, attaining 0.74, as well as improved NDCG to 0.85. As a result, the model is above a high capacity to correctly sort videos regarding their relevance for the user's search. Hence, the results obtained in this study confirm the proposed model's reliability and scalability, making it a relevant tool for educational platforms that seek to satisfy their users' needs through providing relevant and tailored content in an efficient manner. Thus, the framework's unified structure, which couples transcription, semantic matching, classification, and ranking, allows for not only a more accurate and efficient process but also a 25% decrease in computational costs and 20% inferences time, effectively positioning the model as highly applicable for real-time applications, where rapid response and processing are a priority.

Looking ahead, there are several promising research and development scopes. One area of potential inquiry is related to adapting the proposed framework for handling multiple language datasets & samples. This could further enhance the universality of the approach, allowing it to be used in different educational contexts internationally. For instance, the approach could be substantially improved by fine-tuning the performance of the models presented here on specific multilingual corpora. Besides, it could be essential to adapt the pre-trained MMF and RankNet models to take into consideration the specifics of different languages. Alternatively, improved approaches to the semantic analysis in the search could be applied by utilizing other state-of-the-art natural language understanding models like various types of Transformers. Such models are characterized by having deeper contextual embeddings, which could potentially allow improving the performance in the semantic pairing and classification tasks, and thus, achieving even better accuracy rates. Another promising research direction could be related to the integration of user feedback loops, which could be used to dynamically adjust the performance of the ranking algorithms. Following that, the recommendations will become more personalized based on the current needs and preferences of the users.

REFERENCES

[1] W. Jo et al., "Simultaneous Video Retrieval and Alignment", in IEEE Access, vol. 11, pp. 28466-28478, 2023, doi: 10.1109/ACCESS.2023.3259733.

[2] L. Vadicamo et al., "Evaluating Performance and Trends in Interactive Video Retrieval: Insights From the 12th VBS Competition", in IEEE Access, vol. 12, pp. 79342-79366, 2024, doi: 10.1109/ACCESS.2024.3405638.

[3] P. Xu et al., "Fine-Grained Instance-Level Sketch-Based Video Retrieval", in IEEE Transactions on Circuits and Systems for Video Technology, vol. 31, no. 5, pp. 1995-2007, May 2021, doi: 10.1109/TCSVT.2020.3014491.

[4] H. Yoon and J. -H. Han, "Content-Based Video Retrieval With Prototypes of Deep Features", in IEEE Access, vol. 10, pp. 30730-30742, 2022, doi: 10.1109/ACCESS.2022.3160214.

[5] G. Ren, X. Lu and Y. Li, "Joint Face Retrieval System Based On a New Quadruplet Network in Videos of Multiple Camera", in IEEE Access, vol. 9, pp. 56709-56725, 2021, doi: 10.1109/ACCESS.2021.3072055.

[6] S. R. Dubey, "A Decade Survey of Content Based Image Retrieval Using Deep Learning", in IEEE Transactions on Circuits and Systems for Video Technology, vol. 32, no. 5, pp. 2687-2704, May 2022, doi: 10.1109/TCSVT.2021.3080920.

[7] H. Kou, Y. Yang and Y. Hua, "KnowER: Knowledge enhancement for efficient text-video retrieval", in Intelligent and Converged Networks, vol. 4, no. 2, pp. 93-105, June 2023, doi: 10.23919/ICN.2023.0009.

[8] L. Rossetto et al., "Interactive Video Retrieval in the Age of Deep Learning – Detailed Evaluation of VBS 2019", in IEEE Transactions on Multimedia, vol. 23, pp. 243-256, 2021, doi: 10.1109/TMM.2020.2980944.

[9] R. Zuo et al., "Fine-Grained Video Retrieval With Scene Sketches", in IEEE Transactions on Image Processing, vol. 32, pp. 3136-3149, 2023, doi: 10.1109/TIP.2023.3278474.

[10] P. Maniotis and N. Thomos, "Tile-Based Edge Caching for 360° Live Video Streaming", in IEEE Transactions on Circuits and Systems for Video Technology, vol. 31, no. 12, pp. 4938-4950, Dec. 2021, doi:

10.1109/TCSVT.2021.3055985.

[11] W. Jo, G. Lim, J. Kim, J. Yun and Y. Choi, "Exploring the Temporal Cues to Enhance Video Retrieval on Standardized CDVA", in IEEE Access, vol. 10, pp. 38973-38981, 2022, doi: 10.1109/ACCESS.2022.3165177.

[12] D. Han, X. Cheng, N. Guo, X. Ye, B. Rainer and P. Priller, "Momentum Cross-Modal Contrastive Learning for Video Moment Retrieval", in IEEE Transactions on Circuits and Systems for Video Technology, vol. 34, no. 7, pp. 5977-5994, July 2024, doi: 10.1109/TCSVT.2023.3344097.

[13] H. Tang, J. Zhu, M. Liu, Z. Gao and Z. Cheng, "Frame-Wise Cross-Modal Matching for Video Moment Retrieval", in IEEE Transactions on Multimedia, vol. 24, pp. 1338-1349, 2022, doi: 10.1109/TMM.2021.3063631.

[14] H. Sun, J. Xu, J. Wang, Q. Qi, C. Ge and J. Liao, "DLI-Net: Dual Local Interaction Network for Fine-Grained Sketch-Based Image Retrieval", in IEEE Transactions on Circuits and Systems for Video Technology, vol. 32, no. 10, pp. 7177-7189, Oct. 2022, doi: 10.1109/TCSVT.2022.3171972.

[15] J. P. Ebenezer, Z. Shang, Y. Wu, H. Wei, S. Sethuraman and A. C. Bovik, "ChipQA: No-Reference Video Quality Prediction via Space-Time Chips", in IEEE Transactions on Image Processing, vol. 30, pp. 8059-8074, 2021, doi: 10.1109/TIP.2021.3112055.

[16] F. Liu et al., "Infrared and Visible Cross-Modal Image Retrieval Through Shared Features", in IEEE Transactions on Circuits and Systems for Video Technology, vol. 31, no. 11, pp. 4485-4496, Nov. 2021, doi: 10.1109/TCSVT.2020.3048945.

[17] H. Fang, P. Xiong, L. Xu and W. Luo, "Transferring Image-CLIP to Video-Text Retrieval via Temporal Relations", in IEEE Transactions on Multimedia, vol. 25, pp. 7772-7785, 2023, doi: 10.1109/TMM.2022.3227416.

[18] J. Dong, X. Wang, L. Zhang, C. Xu, G. Yang and X. Li, "Feature Re-Learning with Data Augmentation for Video Relevance Prediction", in IEEE Transactions on Knowledge and Data Engineering, vol. 33, no. 5, pp. 1946-1959, 1 May 2021, doi: 10.1109/TKDE.2019.2947442.

[19] Z. Zhang et al., "Chinese Title Generation for Short Videos: Dataset, Metric and Algorithm", in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 46, no. 7, pp. 5192-5208, July 2024, doi: 10.1109/TPAMI.2024.3365739.

[20] [20] N. A. Nasir and S. -H. Jeong, "Fast Content Delivery Using a Testbed-Based Information-Centric Network", in IEEE Access, vol. 9, pp. 101600-101613, 2021, doi: 10.1109/ACCESS.2021.3096042.

[21] F. Zhang, M. Xu and C. Xu, "Geometry Sensitive Cross-Modal Reasoning for Composed Query Based Image Retrieval", in IEEE Transactions on Image Processing, vol. 31, pp. 1000-1011, 2022, doi: 10.1109/TIP.2021.3138302.

[22] Y. Zhang, Q. Qian, H. Wang, C. Liu, W. Chen and F. Wang, "Graph Convolution Based Efficient Re-Ranking for Visual Retrieval", in IEEE Transactions on Multimedia, vol. 26, pp. 1089-1101, 2024, doi: 10.1109/TMM.2023.3276167.

[23] B. Yang, M. Cao and Y. Zou, "Concept-Aware Video Captioning: Describing Videos With Effective Prior Information", in IEEE Transactions on Image Processing, vol. 32, pp. 5366-5378, 2023, doi: 10.1109/TIP.2023.3307969.

[24] O. Tursun, S. Denman, S. Sivapalan, S. Sridharan, C. Fookes and S. Mau, "Component-Based Attention for Large-Scale Trademark Retrieval", in IEEE Transactions on Information Forensics and Security, vol. 17, pp. 2350-2363, 2022, doi: 10.1109/TIFS.2019.2959921.

[25] H. Zhang, A. Sun, W. Jing, L. Zhen, J. T. Zhou and R. S. M. Goh, "Natural Language Video Localization: A Revisit in Span-Based Question Answering Framework", in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 44, no. 8, pp. 4252-4266, 1 Aug. 2022, doi: 10.1109/TPAMI.2021.3060449.

..