Computers, Materials & Continua

Tech Science Press

Doi:10.32604/cmc.2025.061702

ARTICLE

Multi-Label Movie Genre Classification with Attention Mechanism on Movie Plots

Faheem Shaukat[1], Naveed Ejaz[1,2], Rashid Kamal[3,4], Tamim Alkhalifah[5,*], Sheraz Aslam[6,7,*] and Mu Mu[4]

[1]Department of Computing and Technology, Iqra University, H-9 Campus, Islamabad, 04436, Pakistan
[2]School of Computing, Queens University, Kingston, ON K7L2N8, Canada
[3]School of Computing, Ulster University, Belfast, BT15 1ED, Northern Ireland, UK
[4]Faculty of Arts, Science and Technology, University of Northampton, Waterside Campus, Northampton Northamptonshire, NN1 5PH, UK
[5]Department of Computer Engineering, College of Computer, Qassim University, Buraydah, Saudi Arabia
[6]Department of Computer Science, CTL Eurocollege, Limassol, 3077, Cyprus
[7]Department of Electrical Engineering, Computer Engineering, and Informatics, Cyprus University of Technology, Limassol, 3036, Cyprus

*Corresponding Authors: Tamim Alkhalifah. Email: tkhliefh@qu.edu.sa; Sheraz Aslam. Email: sheraz.aslam@cut.ac.cy

Received: 01 December 2024; Accepted: 25 March 2025; Published: 19 May 2025

ABSTRACT: Automated and accurate movie genre classification is crucial for content organization, recommendation systems, and audience targeting in the film industry. Although most existing approaches focus on audiovisual features such as trailers and posters, the text-based classification remains underexplored despite its accessibility and semantic richness. This paper introduces the Genre Attention Model (GAM), a deep learning architecture that integrates transformer models with a hierarchical attention mechanism to extract and leverage contextual information from movie plots for multi-label genre classification. In order to assess its effectiveness, we assess multiple transformer-based models, including Bidirectional Encoder Representations from Transformers (BERT), A Lite BERT (ALBERT), Distilled BERT (DistilBERT), Robustly Optimized BERT Pretraining Approach (RoBERTa), Efficiently Learning an Encoder that Classifies Token Replacements Accurately (ELECTRA), eXtreme Learning Network (XLNet) and Decoding-enhanced BERT with Disentangled Attention (DeBERTa). Experimental results demonstrate the superior performance of DeBERTa-based GAM, which employs a two-tier hierarchical attention mechanism: word-level attention highlights key terms, while sentence-level attention captures critical narrative segments, ensuring a refined and interpretable representation of movie plots. Evaluated on three benchmark datasets Trailers12K, Large Movie Trailer Dataset-9 (LMTD-9), and MovieLens37K. GAM achieves micro-average precision scores of 83.63%, 83.32%, and 83.34%, respectively, surpassing state-of-the-art models. Additionally, GAM is computationally efficient, requiring just 6.10 Giga Floating Point Operations Per Second (GFLOPS), making it a scalable and cost-effective solution. These results highlight the growing potential of text-based deep learning models in genre classification and GAM's effectiveness in improving predictive accuracy while maintaining computational efficiency. With its robust performance, GAM offers a versatile and scalable framework for content recommendation, film indexing, and media analytics, providing an interpretable alternative to traditional audiovisual-based classification techniques.

KEYWORDS: Multi-label classification; artificial intelligence; movie genre classification; hierarchical attention mechanisms; natural language processing; content recommendation; text-based genre classification; explainable AI (Artificial Intelligence); transformer models; BERT

1 Introduction

Automated classification of movies into genres is a critical task in film data analysis, serving as the foundation for effective content organization, recommendation systems, and audience-specific targeting [1–3]. As the volume of available films continues to expand, manual genre categorization, traditionally performed by studios or distributors, is no longer sufficient to meet the growing demands for precise and scalable genre classification [4]. Automated genre classification enables efficient navigation of large-scale cinematic datasets, providing tailored content recommendations and enhancing user experience in digital media platforms. Using trailers, posters, and metadata works well but needs a lot of computing power and rich audiovisual data. This data is not always available, especially for new movies. A text-only method, using plot descriptions, is a cheaper and easier option. Plot summaries come before trailers or posters, making text-based methods useful for early genre prediction. Despite advancements in multimodal genre classification using audio-visual data, recent developments in Natural Language Processing (NLP) open new avenues for text-based genre classification, specifically through the analysis of plot summaries and synopses, which contain semantically rich information crucial for genre identification [5].

Existing approaches to genre classification utilize audio-visual data from trailers, posters, and promotional media to capture genre-specific elements through visual and auditory features. Movie trailers, rich in both visual and audio cues, are analyzed using deep visual feature extraction, spatiotemporal modelling, and audio-based techniques that leverage sound effects, music, and voice modulation as genre indicators [6–9]. Similarly, movie posters convey genre cues through colour schemes, object compositions, and thematic visuals, with methods ranging from low-level colour analysis to high-level semantic attribute detection [10–13]. Recent advancements involve multimodal approaches that combine audio, visual, and textual features such as audio tracks, poster images, frame sequences, and plot summaries to improve classification accuracy by capturing genre-specific nuances comprehensively [14–16].

Advancements in NLP and neural networks have enabled textual approaches, particularly those leveraging movie plots and synopses, to play a significant role in genre classification. These textual summaries, often available from online databases and Application Programming Interface (APIs) (e.g., Internet Movie Database (IMDb), The Movie Database (TMDb)), carry rich semantic and contextual information about genre affiliation. Effectively capturing this information requires sophisticated models capable of navigating the complexities of natural language and the subtle nuances that distinguish genres. Existing research has explored various textual sources, including subtitles [17], synopses [18], plot summaries [19], movie scripts [20], viewer reviews [21], and social media data [22]. Despite extensive research, the field still faces notable gaps, including:

- Existing methods primarily depend on traditional feature extraction approaches, such as Term Frequency-Inverse Document Frequency (TF-IDF), Bag of Words (BoW), and pre-trained word embeddings. The potential of *transformers and hierarchical attention mechanisms remains largely untapped* in the context of movie genre classification.
- State-of-the-art language models, particularly *DeBERTa*, are underexplored despite their enhanced capacity for contextual understanding, which could greatly benefit genre classification tasks.
- Most current studies lack comprehensive validation of model performance across *multiple benchmark datasets*, raising concerns about existing models' generalizability and robustness.

This paper proposes a novel model, the GAM, which combines hierarchical attention mechanisms with the DeBERTa [23] language model. Hierarchical attention has been used in NLP tasks like document classification and sentiment analysis, but applying it to movie genre classification is new and challenging [24]. Movie plots need a deep understanding of story structures and genre details, which current methods struggle

with. GAM combines hierarchical attention with the DeBERTa model to focus on key narrative elements, like important words or sentences. This makes it well-suited for analyzing movie plots and a unique contribution to the field. The hierarchical attention mechanism allows GAM to focus on the most salient words and sentences for genre classification. At the same time, DeBERTa enhances its ability to capture deep contextual and semantic relationships within the text. By integrating these components, GAM addresses the limitations of prior approaches and achieves superior performance in multi-label movie genre classification. To identify the optimal transformer model, we experimented with several models, including ALBERT [25], BERT [26], DistilBERT [27], RoBERTa [28], ELECTRA [29], XLNet [30], and DeBERTa [23], selecting DeBERTa as the best-performing model with the hierarchical attention mechanism.

The GAM model has been extensively evaluated on three benchmark datasets: Trailers12K [6], LMTD-9 [7], and MovieLens37K [31]. The model has been trained on the Trailers12K dataset and tested on all three datasets to assess its generalization and performance on unseen data. The results have been evaluated using recall, precision, F1-score, and Area Under the Precision-Recall Curve (AUPRC) to compare overall performance with state-of-the-art methods. Additionally, the results have been benchmarked against fine-tuned transformer-based and traditional Machine Learning (ML) models. GAM has achieved micro, macro, weighted, and sample average precision scores of 83.63%, 82.16%, 81.73%, and 82.16% on Trailers12K; 83.32%, 82.71%, 81.66%, and 82.71% on LMTD-9; and 83.34%, 82.11%, 81.77%, and 82.11% on MovieLens37K, demonstrating superior performance compared to existing methods.

The primary contributions of this research are as follows:

i. **Proposal of the Genre Attention Model (GAM):** This work introduces GAM, a novel model integrating the DeBERTa transformer with a hierarchical attention mechanism to capture complex, genre-specific details within movie plot text. It addresses limitations in prior genre classification models by enhancing interpretability and contextual understanding of narrative details.
ii. **Development of an Enhanced Attention Mechanism:** GAM incorporates a hierarchical attention mechanism that dynamically prioritizes keywords and sentences, enabling precise genre predictions. This mechanism emphasizes the most salient textual components, significantly improving classification accuracy.
iii. **Demonstration of Superior Classification Accuracy:** xtensive experimentation on the Trailers12K, LMTD-9, and MovieLens37K datasets shows GAM achieving over 83% in precision metrics, outperforming existing state-of-the-art models by up to 8%, demonstrating its efficacy in multi-label genre classification.
iv. **Comprehensive Benchmarking against Transformer-Based Models:** GAM is rigorously compared with high-performing transformer models, including BERT, RoBERTa, and XLNet. Results highlight the synergy between rich language representations, such as DeBERTa, and hierarchical attention mechanisms, emphasizing their combined impact on genre classification accuracy.

The paper is organized as follows: Section 2 reviews multi-label genre classification. Section 3 describes the GAM architecture, detailing the input, DeBERTa, attention, dense, dropout, and output layers. Section 4 presents the experimental setup, results on the Trailers12K, LMTD-9, and MovieLens37K datasets, and comparisons with state-of-the-art methods, including model interpretability through attention weights. Section 5 discusses the model's strengths, limitations, and FLOPS (Floating Point Operations Per Second) efficiency. Section 6 concludes with key findings and future research directions.

2 Related Work

This section provides a systematic review of current multi-label movie genre classification methods, focusing on approaches using trailers, posters, textual data, and multimodal strategies. The methodologies are categorized by their primary modality, while also discussing commonly used datasets.

2.1 Trailer-Based Approaches

Movie trailers serve as a condensed representation of a film's narrative, visual style, and audio cues, making them a valuable resource for genre classification. Previous studies have leveraged various deep-learning methods to extract relevant features. Wehrmann et al. [32] utilized deep visual features from trailer frames, employing a 152-layer ConvNet (Convolutional Network) to generate temporal representations for classification. Bi et al. [33] introduced a Convolutional 3D + Long Short-Term Memory (C3D-LSTM) model to capture spatiotemporal features, while Lin et al. [34] combined visual and audio elements for enhanced genre recognition. TimeSformer, proposed by Bertasius et al. [35], captures temporal relationships across video frames, and though not explicitly applied to genre classification, it demonstrates the potential for this purpose.

Audio features are also critical in trailer analysis. Sharma et al. [36] extracted audio characteristics for genre prediction, and Bhattacharjee et al. [8] employed an Attention Convolutional Neural Network (ACNN) utilizing speech-music confidence sequences to boost classification accuracy. Studies combining multiple features, such as Montalvo et al. [6] and Shambharkar et al. [9], demonstrated the benefits of integrating audio-visual data with 3D Convolutional Neural Networks for enriched spatial and temporal information.

2.2 Poster-Based Approaches

Movie posters are powerful visual representations, often containing genre-specific cues through design elements like color schemes, objects, and layout. Barney et al. [10] analyzed RGB (Red, Green, Blue) color distributions in posters to build predictive feature matrices, while Narawade et al. [11] employed dominant colors and Global Image Descriptor (GIST) descriptors to classify visual content. More advanced techniques incorporate object detection and semantic feature extraction; Chu et al. [37] used Convolutional Neural Network (CNNs) with YOLO (You Only Look Once) object detection model to detect genre-relevant objects, while Wi et al. [12] categorized posters by object themes, enhancing genre prediction for action and horror films. High-level attributes like aesthetics, emotion, and typography were explored by Popat et al. [38], who applied colour theory and design principles to convey genre associations, and Sirattanajakarin et al. [39], who annotated posters with twelve semantic features for genre classification.

2.3 Multimodal Approaches

Integrating multiple modalities-such as visual, textual, and audio, features more comprehensive genre classification models [40]. Nambiar et al. [41] combined visual features from posters (using Visual Geometry Group (VGG16) and Residual Networks (ResNet50)) with textual data (Word2Vectors and GlobalVectors embeddings) for improved classification. Similarly, Braz et al. [42] fused text from movie synopses with poster images processed through Densely Connected Convolutional Networks (DenseNet-169) in a multimodal fusion module. More complex multimodal approaches include Paulino et al. [43], who integrated visual, textual, and audio data using an Inflated 3D (I3D) model, and Mangolin et al. [15], who combined audio, video frames, and textual features via late fusion strategies.

Some studies incorporated metadata for enhanced classification. For instance, Kerger et al. [44] utilized movie summaries and metadata attributes like production year, budget, and company in training models.

Liang et al. [45] extracted features from individual modalities, combining them through global average pooling for effective fusion. These multimodal approaches leverage broader data sources to capture genre nuances more effectively.

2.4 Textual Approaches

Textual data such as plot summaries, synopses, and subtitles provide rich, narrative-based information for genre classification. Portolese et al. [46] employed TF-IDF and pre-trained word embeddings to extract features from synopses, while other studies [18,47] applied vectorization techniques with machine learning algorithms (e.g., CountVectorizer (CV), K-Nearest Neighbors (KNN), Naive Bayes (NB), and Support Vector Machine (SVM)). Deep Learning (DL) approaches, including Bidirectional-Long Short Term Memory (Bi-LSTM) networks [48] and models incorporating the Universal Sentence Encoder (USE) [19], have also been used to capture genre-relevant text patterns. Topic modelling has shown promise, with Matthews et al. [49] demonstrating that topic-based features enhance genre prediction accuracy.

Scripts and subtitles offer additional avenues for genre classification. Agarwal et al. [20] used character interactions within movie scripts, represented through eigenvectors of Laplacian matrices, to predict genres. Hasan et al. [17] utilized TF-IDF and BoW representations to extract features from subtitles, while Rajput et al. [50] used high-frequency words in subtitles as genre indicators. Textual approaches are also adopted in some other studies like [51,52].

Text-based methods lag behind multimodal models, which capture extra details from trailers, posters, and audio. However, text-based methods are still useful when visual data is missing or hard to process, especially for older or independent films. This shows the importance of advanced text models like GAM, which can work well using only plot descriptions.

2.5 Datasets for Genre Classification

Several datasets have facilitated research in movie genre classification. The MM-IMDb dataset [53] contains 26,000 movies annotated with multiple genre labels and includes textual, visual, and auditory data. The MovieLens dataset [31] is widely used in recommendation systems and includes 20 million ratings and genre labels for 27,000 movies. MovieScope [54] offers a multimodal dataset with approximately 45,000 movies, including descriptions, posters, and trailers. The LMDT dataset [7] integrates text descriptions with trailers, covering around 25,000 movies with multiple genre labels, and is available in various versions like LMDT-9 and LMDT-4, focusing on nine and four primary genres, respectively. The Trailers12K dataset [6] includes 12,000 trailers linked to YouTube videos and poster representations, encompassing ten genre categories.

3 Methodology

This research introduces the Genre Attention Model, a novel deep-learning architecture designed for the challenging task of multi-label movie genre classification. GAM tackles the complexity of movie genre classification by combining the strengths of transformer-based language models with a hierarchical attention mechanism, enabling it to discern intricate relationships between a movie's plot description and its associated genres effectively.

3.1 Dataset and Problem Formalization

Movie genre classification is a multi-label classification problem, where a single film can belong to multiple categories simultaneously. Let $D = \{(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)\}$ be the dataset of movies, where N represents the total number of movie samples. Each data point comprises:

- $\mathbf{x}_i$: The textual plot description of the i-th movie, capturing the narrative and key events.
- $\mathbf{y}_i$: A binary vector of fixed length, $y_i \in \{0,1\}^g$, where g is the total number of predefined genres. Each element, y_{ij}, within this vector signals the presence (1) or absence (0) of the j-th genre in the i-th movie. Formally see Eq. (1):

$$y_{ij} = \begin{cases} 1 & \text{if the } j\text{-th genre applies to the } i\text{-th movie,} \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

3.2 Exploration of Transformer-Based Language Models

In exploring transformer-based language models to identify one suited for hierarchical attention mechanisms, we evaluated several advanced architectures, each with unique enhancements to the original BERT framework. ALBERT [25] was designed with parameter reduction techniques that improve memory efficiency, enabling faster training while preserving performance. This efficiency is achieved primarily through factorized embedding parameterization and cross-layer parameter sharing, both of which reduce the model size without significant accuracy loss. BERT [26] itself, the foundation of many subsequent models, introduced the concept of bidirectional training for contextual word representations. This bidirectional approach allows BERT to capture context from both left and right directions simultaneously, a significant advantage in NLP tasks.

DistilBERT is a compressed, version of BERT, optimized for speed and memory efficiency. By retaining approximately 97% of BERT's language understanding ability but with fewer parameters, DistilBERT [27] achieves a balance between performance and computational efficiency, making it suitable for real-time applications. Another refinement of BERT, RoBERTa [28], improved on BERT's training techniques by extending the pretraining phase, using larger batches, removing the Next Sentence Prediction objective, and dynamically adjusting hyperparameters. This rigorous pretraining led to stronger language representation capabilities.

ELECTRA [29] took a novel approach by introducing a discriminator-generator setup during pretraining. Instead of relying on masked language modelling like BERT, ELECTRA's generator replaces words with plausible alternatives, and the discriminator predicts whether each word in the input is correct or replaced. This allows ELECTRA to learn from all tokens rather than just the masked ones, which improves efficiency, particularly for smaller models.

XLNet [30] departed from the masked language modelling objective by using permutation language modelling, a method that combines the benefits of autoregressive and bidirectional language models. By predicting words in a random order, XLNet can capture bidirectional context without masking, which allows it to understand relationships between words in various orders better and avoid limitations in BERT's fixed left-to-right and right-to-left contexts. Mathematically, XLNet maximizes the expected likelihood of a sequence under all possible permutations of factorization orders, which leads to more comprehensive learning of dependencies.

DeBERTa [23] further advanced the BERT framework by incorporating disentangled attention and an improved mask decoder. Disentangled attention in DeBERTa separately encodes the content and position of each word, allowing the model to differentiate between a word's identity and its position more effectively.

This separation leads to a finer-grained understanding of language structure and relationships, crucial for nuanced tasks. The disentangled attention mechanism can be expressed as two distinct functions, $f_{\text{content}}(w)$ and $f_{\text{position}}(p)$, where w represents word embeddings and p denotes position embeddings. By enhancing how DeBERTa processes contextual information, this architecture proved adept at tasks requiring semantic subtlety.

Our experiments indicated that DeBERTa consistently outperformed other models in capturing the semantic nuances within movie plot descriptions, leading to more accurate genre predictions. Given its superior performance, we selected DeBERTa as the core model for our hierarchical attention mechanism.

3.3 Genre Attention Model (GAM) Architecture

GAM seamlessly integrates the DeBERTa language model with a hierarchical attention mechanism, enabling it to focus on the most informative aspects of a movie plot for genre classification. This architecture, visually represented in Fig. 1, comprises the following key components:

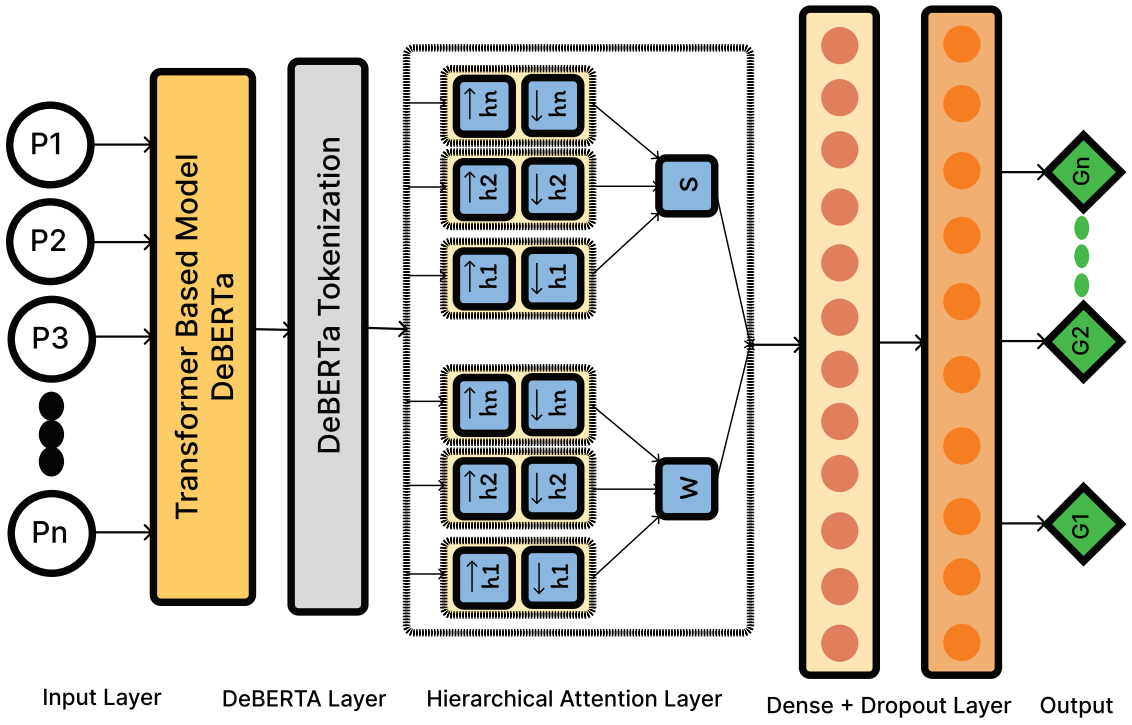

Figure 1: Deep learning model and GAM architecture

3.3.1 Input Layer

The input layer acts as the entry point for the raw plot description, p_n, and prepares it for the model. This description could be anything from a short summary to a full script. It performs crucial preprocessing steps, including:

- Tokenization: The input text is segmented into individual tokens (words or sub-words) using a DeBERTa-specific tokenizer. This tokenizer is responsible for mapping each token to its corresponding numerical ID, creating a sequence of token IDs, $\mathbf{t} = [t_1, t_2, \ldots, t_n]$.

- Attention Mask Generation: An attention mask, $\mathbf{m} = [m_1, m_2, \ldots, m_n]$, is created to guide the model's focus. Each element m_i in the mask is 1 if the corresponding token t_i is a meaningful part of the input and 0 if it's a padding token (added to standardize input lengths).

3.3.2 DeBERTa Layer

The DeBERTa layer uses the pre-trained DeBERTa model to transform the token IDs into contextualized word embeddings, $\mathbf{H} = [h_1, h_2, \ldots, h_n]$, where $h_i \in \mathbb{R}^d$ is the embedding of token t_i. These embeddings capture the meaning of individual words and their semantic significance within the surrounding context.

- Disentangled Attention Mechanism: DeBERTa employs a disentangled attention mechanism, separating content and position-based attention. This allows the model to capture both the semantic relationships between words and their positional information within the sentence. The attention score, e_{ij}, between tokens i and j is computed as see Eq. (2):

$$e_{ij} = (\mathbf{W}_Q \mathbf{h}_i)^\top (\mathbf{W}_K \mathbf{h}_j + \mathbf{a}_{ij}) + b_{ij} \tag{2}$$

  where $\mathbf{W}_Q$ and $\mathbf{W}_K$ are weight matrices for the query and key projections, $\mathbf{a}_{ij}$ is the disentangled position embedding, and b_{ij} is a bias term.
- Enhanced Representations: The DeBERTa layer produces enhanced representations, $\mathbf{H}' = [\mathbf{h}'_1, \mathbf{h}'_2, \ldots, \mathbf{h}'_n]$, by applying the disentangled attention mechanism and subsequent transformations. These representations provide a comprehensive view of the input text, integrating content and position-based information.

3.3.3 Hierarchical Attention Layer

The hierarchical attention layer is the core innovation of GAM, enabling the model to selectively attend to the most salient information within the movie plot. It operates at two levels: first, the word level identifies the most important words in each sentence, capturing the essence of each sentence. Then, the sentence level identifies the most important sentences in the plot, focusing on the key elements of the narrative. This dual-level attention allows the model to understand the plot better and figure out the movie's genres more accurately.

- Word-Level Attention: For each sentence in the plot, this mechanism calculates attention weights, α_{ij}, for each word. Let $\mathbf{H}_s = [h_{s1}, h_{s2}, \ldots, h_{sm}]$ represent the DeBERTa output for the s-th sentence. The attention score, e_{ij}, between words i and j in sentence s is calculated using a compatibility function (e.g., dot product) see Eq. (3):

$$e_{ij} = h_{si}^\top h_{sj} \tag{3}$$

  These scores are then normalized using the softmax function see Eq. (4):

$$\alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{k=1}^{m} \exp(e_{ik})} \tag{4}$$

  This results in a set of attention weights, α_{ij}, that represent the importance of each word j concerning the word i. A context vector, c_i, for each word i is then computed as a weighted sum of the word embeddings see Eq. (5):

$$c_i = \sum_{j=1}^{m} \alpha_{ij} h_{sj} \tag{5}$$

This process effectively highlights the most influential words within each sentence, capturing the essence of their contribution to the overall meaning.

- Sentence-Level Attention: This mechanism operates on the output of the DeBERTa layer, which provides a contextualized $\mathbf{C} = [c_1, c_2, \ldots, c_n]$, representation of the entire plot description. It calculates attention weights, β_s, for each sentence, allowing the model to discern which sentences are most indicative of the movie's genre. The attention score, e_s, for sentence s is calculated as see Eq. (6):

$$e_s = \mathbf{v}_s^\top \tanh(\mathbf{W}_c c_s + \mathbf{b}_c) \tag{6}$$

where $\mathbf{v}_s$ is a weight vector for sentence s, $\mathbf{W}_c$ is a weight matrix, and $\mathbf{b}_c$ is a bias term. These scores are normalized using softmax to obtain the attention weights see Eq. (7):

$$\beta_s = \frac{\exp(e_s)}{\sum_{k=1}^{n} \exp(e_k)} \tag{7}$$

Finally, a comprehensive sentence representation, $\mathbf{r}$, is computed as a weighted sum of the context vectors see Eq. (8):

$$\mathbf{r} = \sum_{s=1}^{n} \beta_s c_s \tag{8}$$

This hierarchical attention mechanism allows GAM to discern not only which words are important within each sentence but also which sentences contribute most significantly to the overall genre classification.

3.3.4 Output Layer

The output layer transforms the aggregated sentence representation, $\mathbf{r}$, into a format suitable for genre prediction. It consists of the following sub-components:

- Dense Layer with ReLU Activation: This layer applies a linear transformation to the sentence representation, followed by the ReLU (Rectified Linear Unit) activation function. ReLU introduces non-linearity into the model, enabling it to learn complex patterns and relationships between the input features and the output genres.
- Dropout Layer: A dropout layer with a rate of 0.5 is incorporated to prevent overfitting. Dropout randomly deactivates a fraction of the neurons during training, forcing the network to learn more robust and generalizable features.
- Dense Layer with Sigmoid Activation: The final layer in the output stage is a dense layer that uses the sigmoid activation function. This function maps the output values to a range between 0 and 1, representing probabilities. Each of the multiple sigmoid outputs from this single dense layer corresponds to a specific genre, signifying the probability that the input movie belongs to that genre.

By combining these layers, the output layer effectively transforms the high-level representation of the movie plot into a set of genre probabilities.

3.4 Prediction and Model Training

A threshold is applied to the output probabilities to obtain the final genre predictions. For instance, if the threshold is set to 0.5, any genre with a probability greater than 0.5 is considered present in the movie. This results in a binary prediction vector for each movie, indicating its predicted genres. Sometimes, a movie might not be assigned any genre if all the predicted probabilities fall below the threshold.

The input to the model is tokenized with a maximum length of 100 tokens. The core of the model is the pre-trained `DeBERTa-base` model from Microsoft, which includes 12 hidden layers, 768 hidden units, and a maximum sequence length of 512. The hierarchical attention layer uses trainable weight matrices, bias vectors, and context vectors to calculate attention scores, determining the importance of different words and sentences in the input plot description. Following the attention layer, a dense layer with 128 neurons and ReLU activation is used. A dropout layer with a rate of 0.5 helps prevent overfitting. The output layer consists of 10 neurons (corresponding to the number of genres) with sigmoid activation, producing probabilities for each genre.

The GAM model undergoes training using the Adam optimizer with a learning rate of 3×10^{-5}. We utilize binary cross-entropy as the loss function, which is suitable for multi-label classification problems, and track accuracy as a metric. Early stopping monitors validation loss with patience set to 3, restoring the best weights if no improvement occurs. Training is conducted over 20 epochs with a batch size of 32.

4 Experiments and Results

This section presents our results on multi-label genre classification for movie trailers, detailing the datasets used (Trailers12K, LMTD-9 and MovieLens37K) and the evaluation metrics applied to assess model performance. We begin by outlining the baseline methods, which include various fine-tuned transformer-based models that serve as benchmarks. We then highlight the performance of our proposed model, GAM, comparing it to these baselines and other state-of-the-art approaches on both datasets. Additionally, we examine the class-specific results of GAM, providing insights into how well it performs across different genres. Finally, we discuss GAMs efficiency in terms of model parameters and FLOPs, assessing its balance between accuracy and computational cost.

4.1 DataSets and Evaluation Metrics

We used three datasets, Trailers12K [6], LMTD-9 [7] and MovieLens37K [31]. The training was conducted using the training sets provided by the Trailers12K dataset. In this section, we provide a brief description of these two datasets. Additionally, we explain the evaluation metrics used in our study.

4.1.1 Trailers12k, LMTD-9 and MovieLens37K Datasets

The Trailers12k, LMTD-9 and MovieLens37K datasets are essential resources for movie trailer analysis aimed at multi-label genre classification, as summarized in Table 1. Trailers12k [6] consists of 12,000 carefully curated movie trailers, each linked with a YouTube trailer, a movie poster, and metadata from IMDb. With verified title-trailer pairs, Trailers12k tags each trailer with up to ten popular genres like action, comedy, drama, and thriller. However, genres like drama (34%) are more common, leading to some imbalance. To address this, it provides stratified splits (70% training, 10% validation, 20% test) that maintain genre balance. Trailers12k also includes frame-level, clip-level, and poster representations, as well as predefined evaluation splits.

LMTD-9 [7] is a focused subset of the larger LMTD dataset, concentrating on nine main genres, excluding fantasy. With 4007 trailers tagged with up to three genres, LMTD-9 has a similar distribution to Trailers12k, with drama (36%) as the most common, followed by comedy (24%) and action (16%). Sci-fi and horror are less frequent (around 5% each). LMTD-9 is divided into training, validation, and test sets, ensuring balanced genre representation.

Table 1: Genre-wise distribution of Trailer12K, LMTD-9, and MovieLens37K datasets across training, test, and validation sets

Genre	Trailer12K[1]			LMTD-9[2]			MovieLens37K[3]		
	Train	Test	Valid	Train	Test	Valid	Train	Test	Valid
Action	2176	623	316	612	163	77	5201	1535	798
Adventure	1366	390	195	433	108	51	3465	947	513
Comedy	2936	839	419	1101	301	147	9747	2757	1363
Crime	1726	497	261	479	121	59	5012	1472	709
Drama	4177	1194	597	1439	390	192	14794	4274	2098
Fantasy	1089	314	165	0	0	0	1705	490	228
Horror	1801	515	257	323	76	32	4027	1127	587
Romance	1529	446	218	470	122	59	4622	1347	651
Sci-Fi	1068	313	154	229	56	25	2026	570	309
Thriller	3106	889	452	501	129	61	7741	2213	1121

[1]Available at https://richardtml.github.io/trailers12k/ (accessed on 24 March 2025); [2]Available at https://github.com/jwehrmann/lmtd (accessed on 24 March 2025); [3]Available at https://grouplens.org/datasets/movielens/ (accessed on 24 March 2025).

The MovieLens37K [31] dataset is a popular choice for research on recommendation systems. It helps analyze movie genres, user preferences, and ratings. In this study, we use a smaller version called MovieLens37K, which focuses on 10 popular genres. This selection allows for a clear analysis of genre classification tasks. The dataset includes 37,000 movies and is split into (70% for training, 10% for validation, and 20% for testing), as shown in Table 1.

Trailers12k, LMTD-9 and MovieLens37K reflect genre popularity trends in movies and offer a rich basis for developing multi-label genre classification models for movie trailers. These datasets are widely used in research and contain movies plot text data. Trailers12K and LMTD-9 use trailers, while MovieLens37K is common in recommendations. Together, they ensure reliable results.

4.1.2 Evaluation Metrics

Four metrics were selected based on the area under the precision-recall curve, commonly employed in multi-label trailer classification studies. μAP is calculated by considering all labels as a binary classification task. It gives an overall idea of the model's performance, with more frequent classes significantly influencing the final score. mAP involves individually computing each class's AUC (Area Under the Curve) and then averaging the results. It provides insights into the model's performance across different classes, regardless of their frequency in the dataset. wAP is similar to mAP as it calculates an AUC for each class but considers the frequency of each class by weighting the average accordingly. This means that more frequent genres have a higher impact on the overall performance. sAP focuses on the performance at the example level. It computes an AUC for each example and then averages the results. Apart from these aggregated metrics, precision, recall, and F1-score have been used to present the classwise results [55].

4.2 Baseline Methods for Comparison

For comparison, we fine-tuned seven transformer-based models: ALBERT, BERT, DistilBERT, RoBERTa, ELECTRA, XLNet, and DeBERTa on the Trailers12K dataset. We also evaluated six traditional

Machine Learning models: Decision Trees, K-Nearest Neighbors, Logistic Regression, Naive Bayes, Random Forest, and Support Vector Machine. Research findings suggest that TF-IDF performs well in text classification tasks [56], so we used TF-IDF to convert movie plots into numerical vectors for effective feature extraction in ML models. Additionally, we tested traditional NLP models, including LSTM, CNN, and a hybrid CNN-LSTM, to compare their performance with both traditional ML methods and deep learning transformers. We replaced the softmax activation function with a sigmoid function for multi-label prediction, allowing each genre label to be predicted independently in transformer-based models. Binary cross-entropy loss was used for each label, with the AdamW optimizer applied for practical training and weight decay. Dropout techniques were used to prevent overfitting, and fine-tuning was performed over three epochs to balance learning and avoid overfitting. To harness the advantages of each model and enhance overall classification performance.

The analysis of fine-tuned models on the Trailer12K dataset in Table 2 reveals detailed differences in performance across various metrics, showcasing which models are better suited for this specific task. DeBERTa consistently stands out with top scores across all primary evaluation metrics, particularly in μAP 80.57 ± 0.28 and mAP 79.35 ± 0.27. This model's stability and accuracy indicate that it captures both broader and specific aspects of the data well, outperforming the other models in delivering robust predictions.

Table 2: Comparison of different fine tuned and ML models over Trailer12K dataset

Models	μAP	mAP	wAP	sAP
Decision trees	65.91 ± 0.30	63.84 ± 0.28	62.92 ± 0.28	63.84 ± 0.28
KNN	59.92 ± 0.06	58.41 ± 0.10	58.17 ± 0.18	58.41 ± 0.10
Logistic regression	72.66 ± 0.03	68.87 ± 0.07	70.29 ± 0.05	68.87 ± 0.07
Naive bayes	68.26 ± 0.09	63.48 ± 0.25	65.83 ± 0.06	63.48 ± 0.25
Random forest	67.19 ± 0.19	63.37 ± 0.20	64.84 ± 0.29	63.37 ± 0.20
SVM	74.00 ± 0.04	71.98 ± 0.10	72.41 ± 0.05	71.98 ± 0.10
LSTM	75.50 ± 0.25	73.20 ± 0.30	73.50 ± 0.20	73.20 ± 0.30
CNN	76.20 ± 0.30	74.00 ± 0.35	74.30 ± 0.25	74.00 ± 0.35
Hybrid CNN-LSTM	77.80 ± 0.35	75.50 ± 0.40	75.80 ± 0.30	75.50 ± 0.40
AlBERT	78.04 ± 0.40	75.51 ± 0.29	75.72 ± 0.29	75.51 ± 0.29
BERT	79.48 ± 0.39	78.04 ± 0.41	77.88 ± 0.31	78.04 ± 0.41
DistilBERT	79.64 ± 0.40	77.35 ± 0.12	77.34 ± 0.17	77.35 ± 0.12
RoBERTa	79.93 ± 0.39	77.58 ± 0.66	77.57 ± 0.48	77.58 ± 0.66
ELECTRA	79.08 ± 0.20	76.84 ± 0.46	76.99 ± 0.27	76.84 ± 0.46
XLNet	80.15 ± 0.39	79.01 ± 0.62	78.55 ± 0.40	79.01 ± 0.62
DeBERTa	**80.57 ± 0.28**	**79.35 ± 0.27**	**79.12 ± 0.23**	**79.35 ± 0.27**

Note: Bold values indicate the best performance for each metric.

XLNet also performs impressively, with μAP and mAP values closely following DeBERTa's scores (80.15 ± 0.39 and 79.01 ± 0.62, respectively). Although slightly behind, XLNet still shows strong generalization across the dataset and maintains high consistency, as evidenced by its competitive scores in the wAP and sAP metrics. This positions XLNet as a reliable model choice, though slightly less precise than DeBERTa in certain areas.

BERT, RoBERTa, and DistilBERT fall into a middle-performance range. RoBERTa (μAP = 79.93 ± 0.39, mAP = 77.58 ± 0.66) slightly edges out BERT and DistilBERT, particularly in wAP and sAP, suggesting it can

effectively capture semantic relationships. BERT's overall performance is relatively close, achieving a μAP of 79.48 ± 0.39 and a mAP of 78.04 ± 0.41. DistilBERT, as a lighter version of BERT, slightly underperforms its larger counterpart but shows acceptable results, which could make it a practical choice when computational efficiency is prioritized over marginally higher accuracy.

ELECTRA and AlBERT score the lowest among the tested transformer models, with AlBERT achieving the lowest mAP at 75.51 ± 0.29. These results suggest that both models may struggle to capture the nuances within the Trailer12K dataset as effectively as other architectures. ELECTRA, although not the lowest, still underperforms compared to BERT-derived models, which may indicate less compatibility with this dataset's specific characteristics.

As expected, the traditional machine learning models generally perform less than the transformer-based models. SVM achieves the highest accuracy among these models with a μAP of 74.00 ± 0.04, while KNN shows the lowest with a μAP of 59.92 ± 0.06. This difference highlights the advantages of transformer models in capturing complex relationships in textual data for movie genre classification.

We also tested traditional NLP models like LSTM, CNN, and a hybrid CNN-LSTM. These models are good at finding patterns in sequences and local text features. However, they struggle to understand long-range connections and deep meanings in movie plots. As shown in Table 2, transformer models performed better than these traditional methods.

In summary, the analysis highlights DeBERTa as the most effective model on the Trailer12K dataset, followed closely by XLNet, making both strong candidates for tasks requiring nuanced understanding. RoBERTa, BERT, and DistilBERT offer moderate performance, while AlBERT and ELECTRA appear less suited to this dataset's requirements. This comparative performance insight could guide model selection for similar datasets or tasks.

4.2.1 Transformer Models Combination with Hierarchical Attention Network

The results in Table 3 from combining the attention network with various transformer models reveal clear distinctions in model performance for genre classification tasks. DeBERTa+Attention emerges as the top-performing combination, significantly outperforming all other pairings across each evaluation metric. With a μAP of 83.63 ± 0.29, mAP of 82.16 ± 0.47, and similarly high scores in wAP and sAP, this combination demonstrates that DeBERTa's rich contextual understanding synergizes well with the hierarchical attention mechanism, capturing nuanced genre features effectively. The strong performance suggests that the DeBERTa+Attention pairing is highly adept at identifying both broad and subtle genre-specific patterns.

Table 3: Results of transformer models with the attention network

Models	μAP	mAP	wAP	sAP
ALBERT+Attention	75.47 ± 1.26	72.87 ± 1.42	73.26 ± 1.25	72.87 ± 1.42
BERT+Attention	77.72 ± 0.12	75.18 ± 0.60	75.38 ± 0.53	75.18 ± 0.60
DistilBERT+Attention	77.22 ± 0.26	74.76 ± 0.60	75.11 ± 0.39	74.76 ± 0.60
RoBERTa+Attention	78.17 ± 0.18	76.02 ± 0.19	76.15 ± 0.23	76.02 ± 0.19
ELECTRA+Attention	77.43 ± 0.38	75.68 ± 0.80	75.65 ± 0.59	75.68 ± 0.80
XLNet+Attention	77.63 ± 0.80	75.25 ± 0.88	75.14 ± 0.85	75.25 ± 0.88
DeBERTa+Attention	**83.63 ± 0.29**	**82.16 ± 0.47**	**81.73 ± 0.26**	**82.16 ± 0.47**

Note: Bold values indicate the best performance for each metric.

Among the other combinations, RoBERTa+Attention achieves the second-best results, with a μAP of 78.17 ± 0.18 and mAP of 76.02 ± 0.19. While this combination falls short of the DeBERTa+Attention's scores, the improvement over other models like BERT+Attention and DistilBERT+Attention suggests that RoBERTa's robust pre-training allows it to leverage the hierarchical structure of the attention network, making it a viable option for tasks that don't require the highest precision of DeBERTa+Attention but still benefit from a balanced performance.

The BERT+Attention and DistilBERT+Attention combinations show moderate success, with μAP values of 77.72 ± 0.12 and 77.22 ± 0.26, respectively. BERT+Attention outperforms DistilBERT+Attention slightly across all metrics, indicating that while the distilled, lighter version of BERT offers computational advantages, it sacrifices some accuracy when used with attention network. This performance gap highlights that the BERT model's full capacity contributes more effectively to the hierarchical attention structure for this classification task.

The results of the ELECTRA+Attention and XLNet+Attention combinations indicate lower efficacy in capturing genre-specific information than other models. ELECTRA+Attention achieves a μAP of 77.43 ± 0.38, and XLNet+Attention scores 77.63 ± 0.80, with both combinations showing higher variability in their performance (as indicated by more significant standard deviations). This may suggest that while ELECTRA and XLNet are competitive, they might lack the specific attributes necessary to fully exploit the attention mechanism, making them less consistent and precise for this genre classification task.

Finally, the ALBERT+Attention combination records the lowest scores across all metrics, with a μAP of 75.47 ± 1.26 and mAP of 72.87 ± 1.42. ALBERT's smaller architecture, while efficient, appears to struggle in supporting the hierarchical structure of the network, potentially due to its limitations in representing complex genre-related subtleties compared to larger models like DeBERTa or RoBERTa.

The analysis shows the DeBERTa+Attention combination as the most effective pairing for genre classification, suggesting that DeBERTa's contextual depth works exceptionally well with a hierarchical attention structure. RoBERTa also shows promising results, while BERT and DistilBERT provide moderate performance. ELECTRA, XLNet, and especially ALBERT exhibit limitations when paired with an attention network, indicating that their architectures may not be as compatible with hierarchical attention for this task. These insights can inform model selection for applications with similar hierarchical attention needs in genre-based classification.

4.2.2 Ablation Study of GAM

To see how hierarchical attention and DeBERTa affect GAM, we did an ablation study in Table 4. We tested three versions of the model: (1) GAM with both DeBERTa and hierarchical attention, (2) GAM with only DeBERTa and no hierarchical attention, and (3) GAM with hierarchical attention but without DeBERTa. The results in Table 4 show that both components help improve performance. The full GAM model, with both DeBERTa and hierarchical attention, gave the best results with a μAP of 83.63%. Removing hierarchical attention dropped the score to 80.57%, showing that it helps capture important genre details. When we used hierarchical attention without DeBERTa, the score fell even more to 77.00% μAP, proving that DeBERTa plays a big role in improving accuracy. These results show that hierarchical attention makes genre classification better, and DeBERTa makes the model even stronger.

Table 4: Impact of hierarchical attention and DeBERTa in GAM performance comparison

Model Variant	μAP	mAP	wAP	sAP
GAM (DeBERTa+Attention)	**83.63 ± 0.29**	**82.16 ± 0.47**	**81.73 ± 0.26**	**82.16 ± 0.47**
GAM only with DeBERTa	80.57 ± 0.28	79.35 ± 0.27	79.12 ± 0.23	79.35 ± 0.27
GAM only with Hierarchical Attention	77.00 ± 0.35	75.50 ± 0.40	75.30 ± 0.38	75.50 ± 0.40

Note: Bold values indicate the best performance for each metric.

4.3 Interpretability through Attention Weights in Movie Genre Classification

To showcase the explainability of Genre Attention Model (GAM), which is a combination of DeBERTa+Attention, we visualize the attention weights for a movie plot of "Committed (2000)". The chosen plot corresponds to a movie classified under the **Romance** and **Drama** genres.

4.3.1 Movie Plot Example

"A young woman goes in search of her midlife crisis suffering husband who left her. The plot of 'Committed' is centred around the story of an intense young woman, played by Graham, whose husband leaves her in order to find himself. She then follows him cross-country and when she catches up with him, complications arise."

4.3.2 Visualization of Attention Weights

The hierarchical attention mechanism and DeBERTa's self-attention layers identify the keywords and phrases in the plot that contribute most to the genre classification. The attention weights with threshold of 0.5 for this plot are visualized in Fig. 2, highlighting the attention Weights of words in sentences with level of importance.

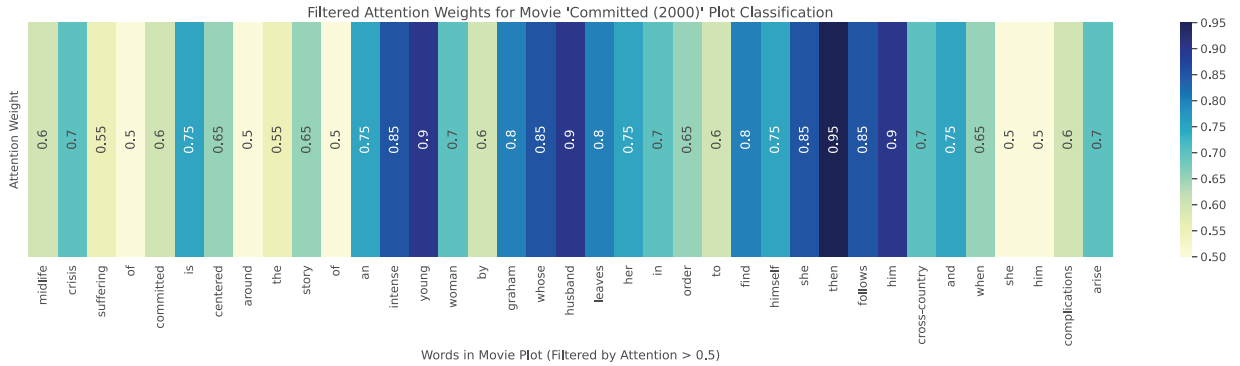

Figure 2: Visualization of GAM model word-level attention weights

4.3.3 Word-Level Attention Weights Table

The attention weights for key words in the movie plot, highlighting their significance for genre classification, are shown in Table 5. These words correspond to key thematic elements, with their interpretations provided there.

Table 5: Word-level attention weights and descriptions for Movie "Committed (2000)" plot classification

Word	Attention weight	Description
Young woman	0.45	Indicates a central character, pivotal for the **Drama** genre, as it sets up the emotional narrative.
Husband leaves her	0.60	Captures the relational conflict, aligning strongly with both **Romance** and **Drama**.
Follows him cross-country	0.55	Suggests themes of pursuit and personal growth, central to **Romance**.
Complications arise	0.50	Adds narrative tension, a hallmark of **Drama**.

4.3.4 Sentence-Level Attention Weight Table

Table 6 shows the average attention weight for each sentence in the movie plot, highlighting the sentences most influential for the genre classification:

Table 6: Sentence-level attention weights for movie "Committed (2000)" plot classification

Sentence	Attention weight
A young woman goes in search of her midlife crisis suffering husband who left her.	0.45
The plot of 'Committed' is centered around the story of an intense young woman, played by Graham, whose husband leaves her in order to find himself.	0.60
She then follows him cross-country and when she catches up with him, complications arise.	0.55

4.3.5 Analysis of Genre Prediction

The model predicted the genres **Romance** and **Drama** with high probabilities of 0.87 and 0.91, respectively. The attention mechanism effectively highlighted key elements that distinguish these genres. Words such as *husband leaves her* and *follows him* received high attention weights for **Romance**, capturing themes of love and emotional pursuit. Similarly, terms like *complications arise* and *intense young woman* strongly aligned with **Drama**, emphasizing conflict and emotional depth inherent to the genre.

4.4 Comparison of GAM with State-of-the-Art Methods on Trailers12K, LMTD9 and MovieLens37K Datasets

In this section, we evaluate the proposed GAM model, which combines DeBERTA+Attention, against state-of-the-art methods on the Trailer12K, LMTD-9 and MovieLens37K datasets.

Trailer12K dataset is divided into three splits, each with different training and test sets. Our comparison, shown in Fig. 3, highlights how GAM performs alongside other notable models like CTT-MMC-A [32], fastVideo [54], TimeSformer [35], and the DIViTA [6] models.

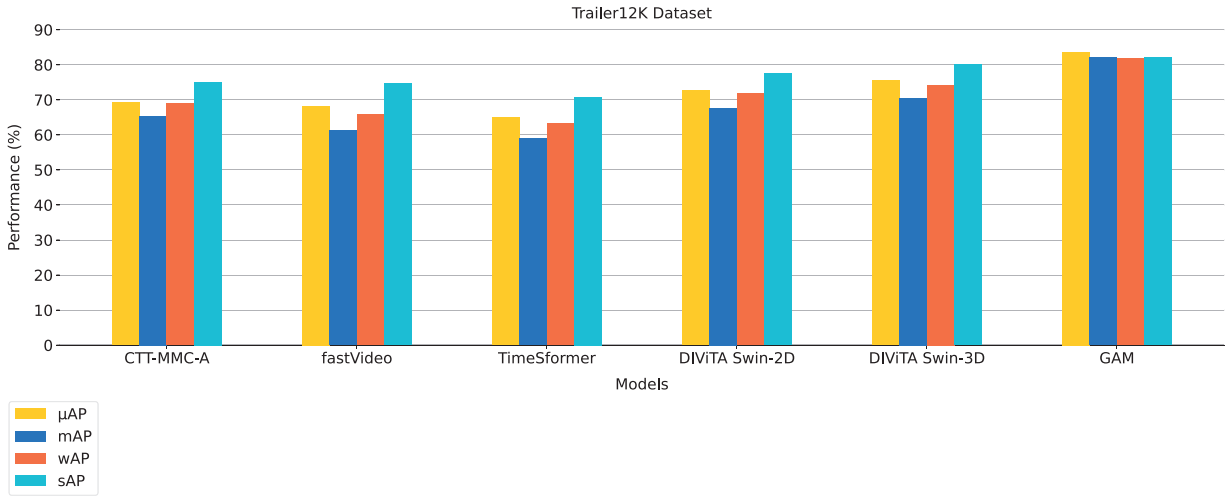

Figure 3: Comparison of GAM with state-of-the-art methods on Trailer12K dataset

GAM scores 83.63% ± 0.29 for μAP, along with 82.16% ± 0.47 for mAP, 81.73% ± 0.26 for wAP, and 82.16% ± 0.47 for sAP. This gives GAM a strong lead over the next best model, DIViTA Swin-3D, which has a μAP of 75.57% ± 0.66 and an mAP of 70.48% ± 0.41. The over 8% difference in μAP clearly shows GAM's ability to capture and generalize the unique features of different genres in trailers.

When we look at the other models, we see that CTT-MMC-A, fastVideo, and TimeSformer don't quite measure up, with μAP scores of 69.27% ± 2.87, 68.21% ± 0.73, and 64.98% ± 1.16, respectively. GAM's strong performance across all metrics showcases its effectiveness and points to its potential for real-world applications where accurate genre classification is vital.

Next, look at the LMTD-9 dataset, which presents a slightly different mix of nine genres. The trained model on the Trailer12K dataset was tested on the LMTD-9 test set, which represents a 20% split of the dataset. Here, GAM continues to impress. According to Fig. 4, it achieves a μAP of 83.32% ± 0.66, mAP of 82.71% ± 0.87, wAP of 81.66% ± 0.78, and sAP of 82.71% ± 0.87. This consistency between both datasets shows that GAM is adaptable and reliable, making it a solid choice for various trailer classification tasks.

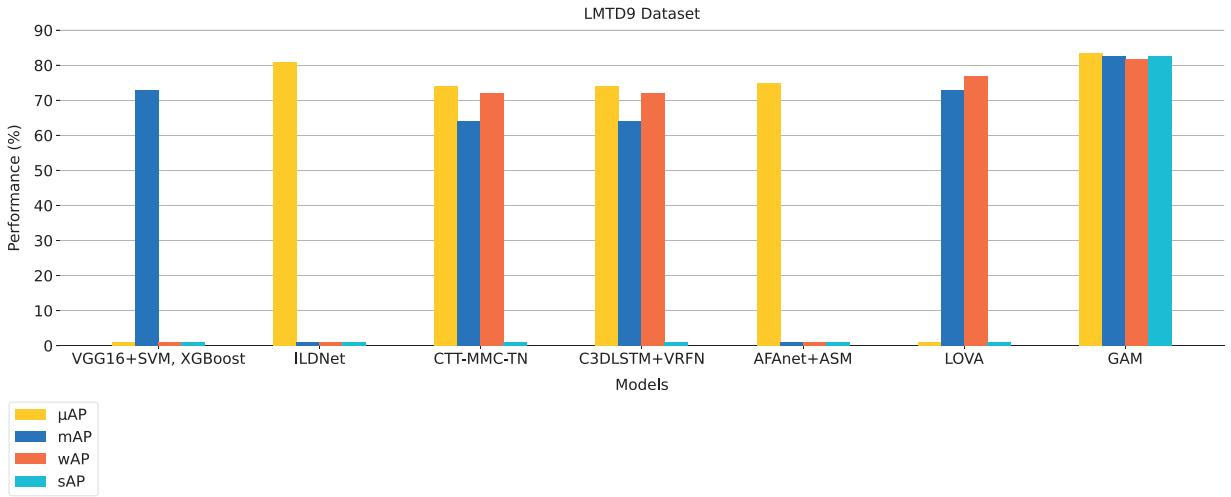

Figure 4: Comparison of GAM with state-of-the-art methods on LMTD9 test dataset

The Trailer12K dataset includes ten genres, while LMTD-9 has nine, with the Fantasy genre excluded from LMTD-9. Regarding competition, ILDNet [57] stands out as the closest contender with a μAP of 81%, but it falls short in providing detailed metrics for a comprehensive comparison. Other models, such as CTT-MMC-TN [32] and AFAnet+ASM [36], score 74% and 75%, respectively, but they don't match the depth and performance that GAM offers. Additionally, VGG16+SVM combined with XGBoost [58] reports a mAP of 73%, though the lack of a μAP score makes it hard to gauge its full effectiveness. C3DLSTM+VRFN [33] achieves a μAP of 74% and an mAP of 64%, showing some capability but also leaving a noticeable gap in its performance across metrics. Meanwhile, LOVA [14] posts an mAP of 73% and a wAP of 77%, but it lacks data for μAP and sAP, which raises concerns about its ability to classify genres accurately.

The results for the MovieLens37K dataset in Fig. 5 show that GAM outperforms all other models in genre classification. It achieves the highest scores across all metrics, with an μAP of 83%, mAP of 80%, wAP of 80%, and sAP of 80%, demonstrating its strong ability to classify genres accurately. In comparison, models like SVM+GD [59] and Bernoulli [60] perform less score but Bi-LSTM+RNN+LR [48] perform effectively high score near to 70% in μAP. DenseNet [61] and ResNet34 [12] also show lower performance, especially with DenseNet achieving just 56% for mAP.

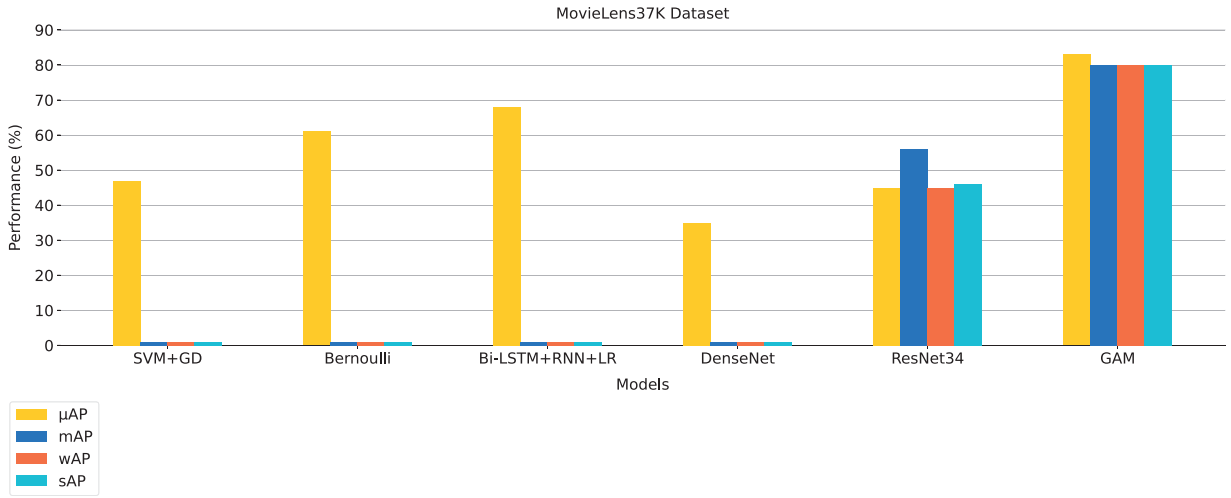

Figure 5: Comparison of GAM with state-of-the-art methods on MovieLens37K test dataset

To sum it all up, GAM stands out as a top performer in the Trailer12K, LMTD-9 and MovieLens37K datasets, showcasing its strengths in multi-label genre classification.

4.4.1 GAM Class wise Results on Trailer12k, LMTD-9 and MovieLens37K

The class-wise performance of GAM on the Trailer12K, LMTD9, and MovieLens37K datasets reveals insightful trends and differences across genres, as shown in Fig. 6. This analysis highlights GAM's strengths and identifies areas for improvement, focusing on precision, recall, and F1 scores for each genre across all three datasets.

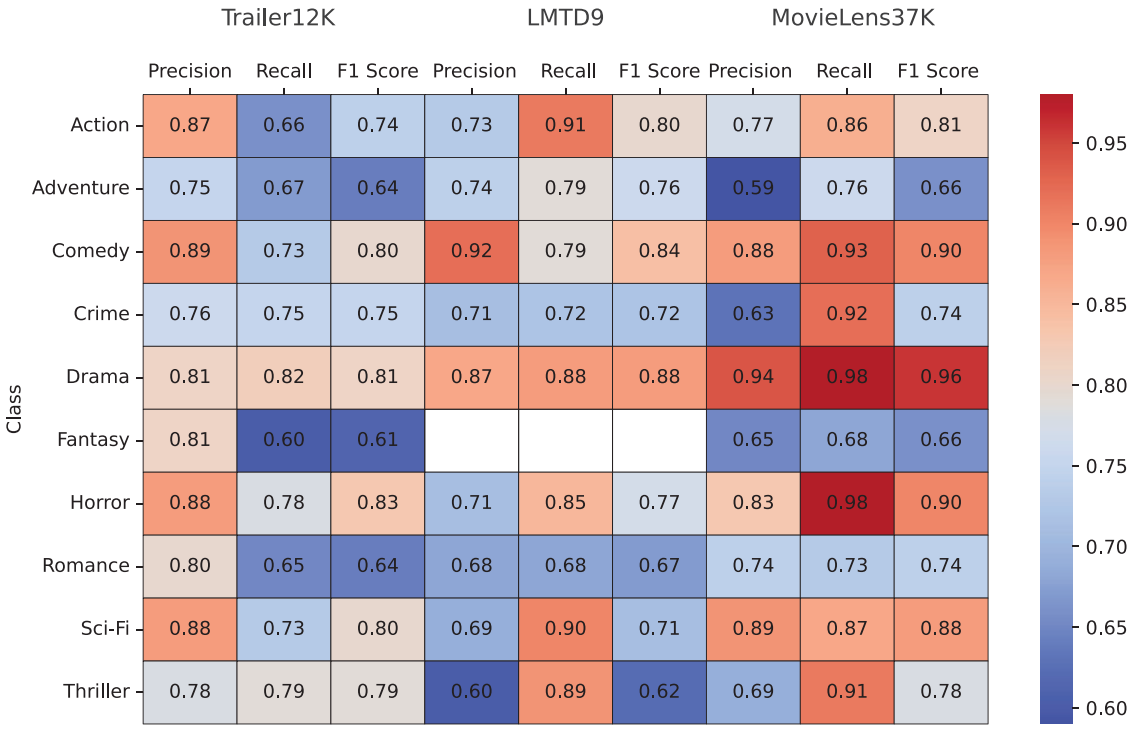

Figure 6: GAM class-wise results for Trailer12K, LMTD9 and MovieLens datasets

GAM performs exceptionally well in identifying genres like *Comedy* and *Drama*. For *Comedy*, the MovieLens37K dataset shows further improvement, with an F1-score of 0.90, surpassing both Trailer12K (0.80) and LMTD9 (0.84). This indicates that MovieLens37K provides rich, diverse samples for comedic content, aiding better recognition. *Drama* achieves the highest F1-score on MovieLens37K at 0.96, compared to 0.81 on Trailer12K and 0.88 on LMTD9, reflecting GAM's robust ability to classify this genre consistently across datasets.

Action exhibits a notable improvement in recall on both LMTD9 (0.91) and MovieLens37K (0.86) compared to Trailer12K (0.66). However, the precision on MovieLens37K (0.77) is slightly lower than Trailer12K (0.87), resulting in a balanced F1-score of 0.81. *Adventure* also sees better performance on MovieLens37K, with an F1-score of 0.66, slightly surpassing Trailer12K (0.64) but lower than LMTD9 (0.76).

Genres like *Romance* and *Fantasy* remain challenging for GAM across all datasets. *Romance* achieves a consistent F1-score of 0.74 on MovieLens37K, a slight improvement over LMTD9 (0.67) but lower than Trailer12K (0.64). *Fantasy* is absent in LMTD9, but on MovieLens37K, it shows a balanced F1-score of 0.66, comparable to Trailer12K (0.61).

For *Horror*, MovieLens37K exhibits the best recall (0.98) among the three datasets, resulting in an F1-score of 0.90, a significant improvement from Trailer12K (0.83) and LMTD9 (0.77). Similarly, *Sci-Fi* shows its strongest performance on MovieLens37K with an F1-score of 0.88, far better than LMTD9 (0.71) and Trailer12K (0.80).

Thriller and *Crime* exhibit mixed trends. On MovieLens37K, *Thriller* achieves a balanced F1-score of 0.78, higher than Trailer12K (0.79) and LMTD9 (0.62). *Crime* remains consistent, with MovieLens37K scoring an F1-score of 0.74, slightly higher than Trailer12K (0.75) and LMTD9 (0.72).

In summary, GAM's performance varies across genres and datasets, with MovieLens37K showing the best results for genres like *Drama*, *Horror*, and *Sci-Fi*. These findings emphasize the importance of dataset diversity and balanced splits in improving model performance for multi-label genre classification tasks. GAM performs well in common genres like Drama and Comedy but struggles with rare ones like Sci-Fi and Horror due to class imbalance. Since movies have multiple genres, oversampling can't fully fix this. Advanced NLP and model improvements can help. Genre overlap adds complexity, but GAM's attention mechanism improves accuracy by focusing on key narrative elements.

4.5 Perfomance Comparison Based on Parameters and FLOPs

Table 7 presents a performance comparison between the GAM and several state-of-the-art models in terms of two key metrics: the number of parameters (in millions) and the floating-point operations per second (FLOPS) measured in gigaflops.

Table 7: Performance comparison of GAM with state-of-the-art methods

Model	Parameters (M)	FLOPS (G)
ALBERT	11.68	28.05
BERT	109.48	27.33
DistilBERT	66.36	13.66
RoBERTa	124.64	26.65
ELECTRA	108.89	27.33
XLNet	116.72	27.96
DeBERTa	138.60	26.66
ALBERT+Attention	12.37	9.00
BERT+Attention	110.17	10.50
DistilBERT+Attention	67.05	8.82
RoBERTa+Attention	125.34	11.50
ELECTRA+Attention	109.58	10.00
XLNet+Attention	117.41	12.00
Bi-LSTM [48]	18.00	7.80
Universal Sentence Encoder (USE) [19]	60.00	12.50
DIViTA Swin-2D [6]	27.90	114.00
DIViTA Swin-3D [6]	27.90	1590.00
GAM	**139.29**	**6.10**

The table lists various models, including ALBERT, BERT, DistilBERT, RoBERTa, ELECTRA, XLNet, and DeBERTa, alongside their corresponding parameters and FLOPS. For instance, BERT, with a parameter count of 109.48 million, exhibits a FLOPS value of 27.33 G. In contrast, ALBERT, which is significantly smaller at 11.68 million parameters, shows a slightly higher FLOPS of 28.05 G, indicating its efficiency despite fewer parameters.

Furthermore, the table highlights the performance of GAM when the attention mechanism is combined with these architectures. For example, the ALBERT+Attention combination features 12.37 million parameters and 9.00 G of FLOPS, demonstrating a reduction in computational load compared to ALBERT alone. Similar patterns are observed with other attention network-enhanced models, such as BERT+Attention and RoBERTa+Attention, which show lower FLOPS values (10.50 and 11.50 G, respectively) despite maintaining a high parameter count.

Moreover, Bi-LSTM (18 M parameters, 7.80 G FLOPS) improves sequence learning with moderate cost, while Universal Sentence Encoder (USE) (60 M parameters, 12.50 G FLOPS) excels in semantic similarity but requires higher computation.

Additionally, the table includes the DIViTA [6] models, specifically the Swin-2D and Swin-3D architectures. These models demonstrate significantly higher FLOPS, particularly the Swin-3D model, which achieves an impressive 1590.00 G but also has a larger parameter count of 27.90 million.

Finally, the performance of GAM is highlighted in bold, showcasing a parameter count of 139.29 million and a notably low FLOPS of 6.10 G, indicating a trade-off between model size and computational efficiency. This comparative analysis emphasizes these models' varying efficiencies and capabilities, shedding light on the potential advantages of integrating GAM with existing architectures.

4.6 True Labels vs. Predicted Labels Analysis for Trailer12K, LMTD-9 and MovieLens37K

The co-occurrence matrices for Trailer12K, LMTD9, and MovieLens37K present a comparative analysis of predicted and true labels across various movie genres, revealing insights into classification performance. These tables provide an evaluation of the model's performance in the multi-label, multi-class setting. Tables 8–10 show the co-occurrence matrices for the Trailer12K, LMTD9, and MovieLens37K datasets, respectively. The tables display the model predictions, with true labels on the vertical axis and predicted labels on the horizontal axis. Diagonal values indicate correct predictions, while off-diagonal values highlight misclassifications.

Table 8: Co-occurrence matrix for Trailer12K

	True labels vs. Predicted labels									
	Action	**Adventure**	**Comedy**	**Crime**	**Drama**	**Fantasy**	**Horror**	**Romance**	**Sci-Fi**	**Thriller**
Action	415	145	50	200	200	57	78	17	129	333
Adventure	181	222	96	21	96	89	32	17	86	69
Comedy	135	113	530	105	340	55	77	166	60	111
Crime	169	18	48	317	306	2	40	8	14	351
Drama	203	91	318	263	830	41	121	191	79	431
Fantasy	103	131	89	9	78	125	65	33	41	56
Horror	68	29	51	28	94	25	368	5	82	319
Romance	31	30	221	38	304	22	18	199	16	53
Sci-Fi	147	83	38	6	60	26	82	6	211	139
Thriller	250	54	58	288	390	20	262	26	125	651

Table 9: Co-occurrence matrix for LMTD9

	True labels vs. Predicted labels								
	Action	**Adventure**	**Comedy**	**Crime**	**Drama**	**Horror**	**Romance**	**Sci-Fi**	**Thriller**
Action	120	31	22	48	28	13	3	35	78
Adventure	58	62	34	4	13	6	2	26	18
Comedy	38	22	221	34	103	27	68	13	46
Crime	53	0	19	80	44	5	3	3	81
Drama	59	18	115	71	283	16	64	9	110

(Continued)

Table 9 (continued)

	True labels vs. Predicted labels								
	Action	**Adventure**	**Comedy**	**Crime**	**Drama**	**Horror**	**Romance**	**Sci-Fi**	**Thriller**
Horror	8	3	8	0	12	60	3	12	37
Romance	4	1	79	8	69	1	68	1	10
Sci-Fi	17	6	5	2	10	20	1	41	17
Thriller	36	1	7	38	46	34	5	13	93

Table 10: Co-occurrence matrix for MovieLens37K

	True labels vs. Predicted labels									
	Action	**Adventure**	**Comedy**	**Crime**	**Drama**	**Fantasy**	**Horror**	**Romance**	**Sci-Fi**	**Thriller**
Action	4301	1454	1534	2248	2606	443	697	276	966	2571
Adventure	1749	1932	997	382	1190	571	279	195	583	722
Comedy	1406	1433	9387	1861	5500	682	982	2781	588	1437
Crime	1254	140	941	3090	2506	32	480	176	68	2516
Drama	2910	2045	5571	3931	15675	932	1603	4685	698	5282
Fantasy	641	866	757	60	686	1090	501	246	341	365
Horror	774	289	969	454	1013	455	3641	87	733	2995
Romance	534	522	3319	669	4540	371	231	3559	151	857
Sci-Fi	1365	785	855	105	639	366	880	120	2017	1157
Thriller	2148	419	1006	3002	3595	250	2346	332	750	5380

From the co-occurrence tables, varying performance between different genres can be observed across the three datasets. For instance, in Table 8, the model exhibits lower accuracy for genres like "Sci-Fi" and "Romance" in the Trailer12K dataset. Similarly, in Table 9, genres like "Action" and "Adventure" show relatively lower accuracy in the LMTD9 dataset. This performance variation can be attributed to several factors, including the imbalanced distribution of samples across different genres in the datasets. Additionally, the semantic and thematic similarities between specific genres, such as "Sci-Fi" and "Action," or "Fantasy" and "Adventure," make it challenging for the model to differentiate between them accurately. Interestingly, Table 10 shows a generally higher accuracy across most genres for the MovieLens37K dataset, suggesting the potential impact of a larger dataset on model performance.

In the Trailer12K (Table 8), the diagonal values represent correctly predicted labels, with "Drama" showing the highest accuracy (830) followed by "Comedy" (530) and "Thriller" (651). The off-diagonal values illustrate misclassifications, notably high for "Action" (415), indicating some overlap between genres, particularly with "Thriller" (333) and "Drama" (200). This overlap is expected, as action movies often incorporate elements of thrill and drama. From the table, the model's difficulty in distinguishing genres like "Sci-Fi" and "Romance" is evident, potentially reflecting the imbalance in data and the semantic overlap between similar genres. For example, a movie labelled as "Action" might also belong to genres like "Thriller" or "Adventure." If the model predicts "Thriller" but not "Action," the co-occurrence matrix will record this as a misclassification for "Action," highlighting the model's potential confusion between the two genres.

For example, a movie labelled as "Action" might also belong to genres like "Thriller" or "Adventure". If the model predicts "Thriller" but not "Action", the co-occurrence matrix will record this as a misclassification for "Action", highlighting the model potential confusion between the two genres.

In contrast, the LMTD9 (Table 9) exhibits a lower overall predictive accuracy. The highest diagonal value is for "Comedy" (221), significantly lower than in Trailer12K. Other genres, like "Drama" (283), also reflect improved prediction compared to others, but overall misclassifications are more pronounced. For instance, "Action" (120) and "Adventure" (62) show much lower correct predictions and higher misclassifications across various genres, with notable confusion in categories such as "Crime" (81) and "Thriller" (93). This lower accuracy, particularly for "Action" and "Adventure," may be attributed to the imbalanced data and the inherent difficulty in distinguishing between genres with thematic similarities, such as "Crime" and "Thriller."

Finally, Table 10 presents the co-occurrence matrix for the MovieLens37K dataset. This table shows a generally higher accuracy across most genres compared to the Trailer12K and LMTD9 datasets. "Drama" exhibits the highest number of correct predictions (15675), followed by "Comedy" (9387) and "Action" (4301). However, similar patterns of misclassification persist, particularly between genres with close thematic relationships. For example, there's notable confusion between "Action" and "Thriller," "Comedy" and "Romance," and "Sci-Fi" and "Thriller." This highlights the ongoing challenge of accurately classifying movies with overlapping genre elements, even with a larger and potentially more balanced dataset.

5 Discussion and Limitations

GAM demonstrates significant advancements in text-based multi-label genre classification by combining the DeBERTa language model with hierarchical attention mechanisms. GAM's text-only method is great for streaming platforms. It improves recommendations and tagging, even for released and unreleased movies with little or no audiovisual data. It also runs efficiently in real-time for audience targeting and content organization. Platforms like Netflix and Amazon Prime can use it to sort movies by genre using plot summaries. GAM handles overlapping genres better than traditional models by focusing on key narrative elements. It links *chase* and *explosion* to Action, while *suspense* signals Thriller. Unlike fixed-rule models, GAM adapts to context, improving accuracy and clarity. This approach capitalizes on the textual richness of plot descriptions, achieving high classification accuracy on both the Trailers12K, LMTD-9 and MovieLens37K datasets. GAM's performance, particularly its μAP scores of 83.63% on Trailers12K, 83.32% on LMTD-9 Tables and 83.34% on MovieLens37K, outperforms current multimodal state-of-the-art models. Notably, GAM's success emphasizes the potential for sophisticated text-only models to serve as reliable tools for genre classification, especially when computational or data limitations make multimodal approaches less feasible.

The model's class-wise results (Fig. 6) reveal GAM's ability to identify key genres effectively, with high F1 scores in *Drama*, *Comedy*, and *Thriller*. This demonstrates that hierarchical attention can capture genre-defining words and phrases within plot summaries, giving GAM a detailed understanding of diverse narratives. Moreover, GAM's efficiency in terms of computational load (Table 7) requiring just 6.10 GFLOPS compared to models like DIViTA Swin-3D-reinforces its suitability for real-world applications, where efficient processing is essential. By prioritizing salient information in plot descriptions, GAM offers a precise and computationally viable solution for genre classification, even for large-scale streaming and recommendation systems.

Despite these strengths, GAM has limitations. Its reliance on plot summaries makes it less effective for genres that heavily rely on audio-visual cues, such as *Action* and *Horror*. Text-based methods miss key details like visuals, sounds, and music, which multimodal models use for better accuracy. Additionally, GAM's independence in handling each genre label might miss co-occurrences and interdependencies between

genres, as shown by the co-occurrence matrices Tables 8–10, where misclassifications are often in closely related genres. Addressing these aspects through multimodal data integration or inter-genre dependency modeling could further refine GAM's classification accuracy and broaden its applicability across various content types.

6 Conclusions

This paper presented the Genre Attention Model (GAM), a novel architecture designed to enhance multi label movie genre classification by integrating hierarchical attention mechanisms with the DeBERTa language model. GAM addresses the challenge of accurately classifying movies into multiple genres using only plot descriptions, focusing on capturing detailed relationships within textual data that differentiate genres. By selectively assigning importance to words and sentences, GAM effectively extracts narrative elements most indicative of specific genres, overcoming limitations in previous methods that often rely on multimodal data and underutilize the semantic richness of textual content. Empirical evaluations on three benchmark datasets Trailers12K, LMTD9, and MovieLens37K demonstrated GAM's superior performance, achieving micro average precision scores of 83.63%, 83.32%, and 83.34%, respectively, surpassing state of the art models. These results highlight GAM's capability to deliver precise genre classification from text alone, streamlining processes by reducing reliance on resource-intensive multimodal data. Its success in narrative-driven genres such as *Drama* and *Romance* further underscores its effectiveness in extracting contextually rich information from plot summaries, making it a practical tool for streaming recommendations and movie tagging, especially when audiovisual data is unavailable or not yet released.

Detailed analysis revealed GAM robustness across diverse genres, achieving high F1 scores in narrative-driven categories but showing limitations in audio visually oriented genres like *Action* and *Horror*. This suggests that while GAM hierarchical attention mechanism excels at text-based genre nuances, some genres may benefit from multimodal inputs that capture non-textual cues, such as visual or auditory elements. Future research could explore integrating additional text-based modalities, such as movie scripts or viewer reviews, to enhance the model's understanding of genre-defining details. Furthermore, incorporating visual and audio features from trailers and posters could improve classification in genres that rely on non-narrative cues. While GAM was designed for movie genre classification, its architecture is adaptable to other domains, such as book genres, news articles, and streaming media recommendations. Refining and expanding the GAM framework could lead to more versatile, domain-adaptable genre classification solutions, leveraging the strengths of hierarchical attention and advanced language modelling.

Acknowledgement: The authors extend their appreciation to the reviewers and editors for their valuable comments and suggestions, which have significantly contributed to improving the quality of this work.

Funding Statement: The researchers would like to thank the Deanship of Graduate Studies and Scientific Research at Qassim University for financial support (QU-APC-2025).

Author Contributions: The authors confirm their contribution to the paper as follows: Faheem Shaukat: Experimentation, Data Collection, Designing, Methodology, Original Writing; Naveed Ejaz: Supervision, Methodology, Original Writing, Data, Validation; Rashid Kamal: Validation, Methodology, Original Writing, Editing; Tamim Alkhalifah: Writing—Review, Editing; Sheraz Aslam: Writing—Review, Editing; Mu Mu: Methodology, Original Writing, Editing, Validation. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: The dataset and code used in this study are publicly available on Zenodo at the following link: https://doi.org/10.5281/zenodo.14906135.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest to report regarding the present study.

References

1. Mehal AS, Meena K, Singh RB, Shambharkar PG. Movie genres and beyond: an analytical survey of classification techniques. In: 2021 5th International Conference on Trends in Electronics and Informatics (ICOEI); Tirunelveli, India: IEEE; 2021. p. 1193–8. doi:10.1109/ICOEI51242.2021.9453021.
2. Lavanya R, Singh U, Tyagi V. A comprehensive survey on movie recommendation systems. In: 2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS); Coimbatore, India: IEEE; 2021. p. 532–6. doi:10.1109/ICAIS50930.2021.9395759.
3. Wang E, Yang Y, Wu J, Liu W, Wang X. An efficient prediction-based user recruitment for mobile crowdsensing. IEEE Transact Mobile Comput. 2018;17(1):16–28. doi:10.1109/TMC.2017.2702613.
4. Shambharkar PG, Anand A, Kumar A. A survey paper on movie trailer genre detection. In: 2020 International Conference on Computing and Data Science (CDS); Stanford, CA, USA: IEEE; 2020. p. 238–44. doi:10.1109/CDS49703.2020.00055.
5. Shi H, Dao SD, Cai J. LLMFormer: large language model for open-vocabulary semantic segmentation. Int J Comput Vis. 2025;133(2):742–59. doi:10.1007/s11263-024-02171-y.
6. Montalvo-Lezama R, Montalvo-Lezama B, Fuentes-Pineda G. Improving transfer learning for movie trailer genre classification using a dual image and video transformer. Informat Process Manag. 2023;60(3):103343. doi:10.1016/j.ipm.2023.103343.
7. Simões GS, Wehrmann J, Barros RC, Ruiz DD. Movie genre classification with convolutional neural networks. In: 2016 International Joint Conference on Neural Networks (IJCNN); Vancouver, BC, Canada: IEEE; 2016. p. 259–66. doi:10.1109/IJCNN.2016.7727207.
8. Bhattacharjee M, Guha P. Exploration of speech and music information for movie genre classification. ACM Transact Multim Comput Commun Appl. 2024;20(8):241. doi:10.1145/3664197.
9. Shambharkar PG, Thakur P, Imadoddin S, Chauhan S, Doja M. Genre classification of movie trailers using 3d convolutional neural networks. In: 2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS); Madurai, India: IEEE; 2020. p. 850–8. doi:10.1109/ICICCS48265.2020.9121148.
10. Barney G, Kaya K. Predicting genre from movie posters. In: Stanford CS 229: Machine learning; Stanford, California, USA: CS230 Deep Learning, Stanford University; 2019. [cited 2025 Feb 1]. https://api.semanticscholar.org/CorpusID:203596980.
11. Narawade V, Potnis A, Ray V, Rathor P. Movie posters classification into multiple genres. ITM Web Conf. 2021;40(1):03048. doi:10.1051/itmconf/20214003048.
12. Wi JA, Jang S, Kim Y. Poster-based multiple movie genre classification using inter-channel features. IEEE Access. 2020;8(20):66615–24. doi:10.1109/ACCESS.2020.2986055.
13. Hossain N, Ahamad MM, Aktar S, Moni MA. Movie genre classification with deep neural network using poster images. In: 2021 International Conference on Information and Communication Technology for Sustainable Development (ICICT4SD); Dhaka, Bangladesh: IEEE; 2021. p. 195–9. doi:10.1109/ICICT4SD50815.2021.9396778.
14. Cai Z, Ding H, Wu J, Xi Y, Wu X, Cui X. Multi-label movie genre classification based on multimodal fusion. Multim Tools Appl. 2024;83(12):36823–40. doi:10.1007/s11042-023-16121-2.
15. Mangolin RB, Pereira RM, Britto AS, Silla CN, Feltrim VD, Bertolini D, et al. A multimodal approach for multi-label movie genre classification. Multim Tools Appl. 2022;81(14):19071–96. doi:10.1007/s11042-020-10086-2.
16. Shao Y, Guo N. Recognizing online video genres using ensemble deep convolutional learning for digital media service management. J Cloud Comput. 2024;13(1):1–17. doi:10.1186/s13677-024-00664-2.
17. Hasan MM, Dip ST, Rahman T, Akter MS, Salehin I. Multilabel movie genre classification from movie subtitle: parameter optimized hybrid classifier. In: 2021 4th International Symposium on Advanced Electrical and Communication Technologies (ISAECT); Khobar, Saudi Arabia: IEEE; 2021. p. 1–6. doi:10.1109/ISAECT53699.2021.9668427.
18. Buslim N, Oh LK, Hardy MA, Wijaya Y. Comparative analysis of KNN, Naïve Bayes and SVM algorithms for movie genres classification based on synopsis. J Tek Inform. 2022;15(2):169–77. doi:10.15408/jti.v15i2.29302.

19. Kumar N, Kumar S, Dev A, Naorem S. Leveraging universal sentence encoder to predict movie genre. In: 2021 7th International Conference on Advanced Computing and Communication Systems (ICACCS); Coimbatore, India; 2021. p. 1013–8. doi:10.1109/ICACCS51430.2021.9441685.
20. Agarwal D, Vijay D, Ayush, Rahul. Genre classification using character networks. In: 2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS); Madurai, India: IEEE; 2021. p. 216–22. doi:10.1109/ICICCS51141.2021.9432303.
21. González F, Torres-Ruiz M, Rivera-Torruco G, Chonona-Hernández L, Quintero R. A natural-language-processing-based method for the clustering and analysis of movie reviews and classification by genre. Mathematics. 2023;11(23):4735. doi:10.3390/math11234735.
22. Govindaswamy KR, Ragunathan S. Genre classification of Telugu and English movie based on the hierarchical attention neural network. Int J Intell Eng Syst. 2021;14(1):54–62. doi:10.22266/ijies2021.0228.06.
23. He P, Liu X, Gao J, Chen W. DeBERTa: decoding-enhanced BERT with disentangled attention. arXiv:2006.03654. 2020. doi:10.48550/arXiv.2006.03654.
24. Yang Z, Yang D, Dyer C, He X, Smola A, Hovy E. Hierarchical attention networks for document classification. In: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies; San Diego, CA, USA; 2016. p. 1480–9. doi:10.18653/v1/N16-1174.
25. Lan Z, Chen M, Goodman S, Gimpel K, Sharma P, Soricut R. ALBERT: a lite BERT for self-supervised learning of language representations. arXiv:1909.11942. 2019. doi:10.48550/arXiv.1909.11942.
26. Devlin J, Chang MW, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. arXiv:1810.04805. 2018. doi:10.48550/arXiv.1810.04805.
27. Sanh V, Debut L, Chaumond J, Wolf T. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv:1910.01108. 2019. doi:10.48550/arXiv.1910.01108.
28. Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, et al. RoBERTa: a robustly optimized BERT pretraining approach. arXiv:1907.11692. 2019. doi:10.48550/arXiv.1907.11692.
29. Clark K. ELECTRA: pre-training text encoders as discriminators rather than generators. arXiv:2003.10555. 2020. doi:10.48550/arXiv.2003.10555.
30. Yang Z. XLNet: generalized autoregressive pretraining for language understanding. arXiv:1906.08237. 2019. doi:10.48550/arXiv.1906.08237.
31. Harper FM, Konstan JA. The movielens datasets: history and context. ACM Transact Interact Intell Syst (TiiS). 2015;5(4):1–19. doi:10.1145/2827872.
32. Wehrmann J, Barros RC. Movie genre classification: a multi-label approach based on convolutions through time. Appl Soft Comput. 2017;61(1):973–82. doi:10.1016/j.asoc.2017.08.029.
33. Bi T, Jarnikov D, Lukkien J. Video representation fusion network for multi-label movie genre classification. In: 2020 25th International Conference on Pattern Recognition (ICPR); Milan, Italy: IEEE; 2021. p. 9386–91. doi:10.1109/ICPR48806.2021.9412480.
34. Lin F, Yuan J, Chen Z, Abiri M. Enhancing multimedia management: cloud-based movie type recognition with hybrid deep learning architecture. J Cloud Comput. 2024;13(1):104. doi:10.1186/s13677-024-00668-y.
35. Bertasius G, Wang H, Torresani L. Is space-time attention all you need for video understanding? arXiv.2102.05095. 2021. doi:10.48550/arXiv.2102.05095.
36. Sharma A, Jindal M, Mittal A, Vishwakarma DK. A unified audio analysis framework for movie genre classification using movie trailers. In: 2021 International Conference on Emerging Smart Computing and Informatics (ESCI); Online: IEEE; 2021. p. 510–5. doi:10.1109/ESCI50559.2021.9396892.
37. Chu WT, Guo HJ. Movie genre classification based on poster images with deep neural networks. In: Proceedings of the Workshop on Multimodal Understanding of Social, Affective and Subjective Attributes; Mountain View, CA, USA; 2017. p. 39–45. doi:10.1145/3132515.3132516.
38. Popat A, Gupta L, Meedinti GN, Perumal B. Movie poster classification using federated learning. Procedia Comput Sci. 2023;218(5):2007–17. doi:10.1016/j.procs.2023.01.177.

39. Sirattanajakarin S, Thusaranon P. Movie genre in multi-label classification using semantic extraction from only movie poster. In: Proceedings of the 7th International Conference on Computer and Communications Management; Bangkok, Thailand; 2019. p. 23–7. doi:10.1145/3348445.3348475.
40. Shi H, Hayat M, Cai J. Unified open-vocabulary dense visual prediction. IEEE Transact Multim. 2024;26:8704–16. doi:10.1109/TMM.2024.3381835.
41. Nambiar G, Roy P, Singh D. Multi modal genre classification of movies. In: 2020 IEEE International Conference for Innovation in Technology (INOCON); Bangalore, India: IEEE; 2020. p. 1–6. doi:10.1109/INOCON50539.2020.9298385.
42. Braz L, Teixeira V, Pedrini H, Dias Z. Image-text integration using a multimodal fusion network module for movie genre classification. In: 11th International Conference of Pattern Recognition Systems (ICPRS 2021); 2021; Curico, Chile: IET; 2021. p. 200–5. doi:10.1049/icp.2021.1456.
43. Paulino MAD, Costa YM, Feltrim VD. Evaluating multimodal strategies for multi-label movie genre classification. In: 2022 29th International Conference on Systems, Signals and Image Processing (IWSSIP); Sofia, Bulgaria: IEEE; 2022. p. 1–4. doi:10.1109/IWSSIP55020.2022.9854451.
44. Kerger D. The impact of explainable ai on low-accuracy models: a practical approach with movie genre prediction. In: 2024 International Conference on Artificial Intelligence, Computer, Data Sciences and Applications (ACDSA); Mahe, Seychelles: IEEE; 2024. p. 1–5. doi:10.1109/ACDSA59508.2024.10467240.
45. Liang X, Fu P, Guo Q, Zheng K, Qian Y. DC-NAS: divide-and-conquer neural architecture search for multi-modal classification. In: Proceedings of the AAAI Conference on Artificial Intelligence; Vancouver, BC, Canada; 2024. Vol. 38, p. 13754–62. doi:10.1609/aaai.v38i12.29281.
46. Portolese G, Feltrin VD. On the use of synopsis-based features for film genre classification. In: Anais do XV Encontro Nacional de Inteligencia Artificial e Computacional. Porto Alegre, Brazil: SBC; 2018. p. 892–902. doi: 10.5753/eniac.2018.4476.
47. Wang J. Using machine learning to identify movie genres through online movie synopses. In: 2020 2nd International Conference on Information Technology and Computer Application (ITCA); Guangzhou, China: IEEE; 2020. p. 1–6. doi:10.1109/ITCA52113.2020.00008.
48. Ertugrul AM, Karagoz P. Movie genre classification from plot summaries using bidirectional LSTM. In: 2018 IEEE 12th International Conference on Semantic Computing (ICSC); Laguna Hills, CA, USA: IEEE; 2018. p. 248–51. doi:10.1109/ICSC.2018.00043.
49. Matthews P, Glitre K. Genre analysis of movies using a topic model of plot summaries. J Associat Inform Sci Technol. 2021;72(12):1511–27. doi:10.1002/asi.24525.
50. Rajput NK, Grover BA. A multi-label movie genre classification scheme based on the movie subtitles. Multim Tools Applicat. 2022;81(22):32469–90. doi:10.1007/s11042-022-12961-6.
51. Wang T, Hou B, Li J, Shi P, Zhang B, Snoussi H. TASTA: text-assisted spatial and temporal attention network for video question answering. Adv Intell Syst. 2023;5(4):2200131. doi:10.1002/aisy.202200131.
52. Chen S, Wang W, Chen X, Zhang M, Lu P, Li X, et al. Enhancing Chinese comprehension and reasoning for large language models: an efficient LoRA fine-tuning and tree of thoughts framework. J Supercomput. 2024;81(1):50. doi:10.1007/s11227-024-06499-7.
53. Arevalo J, Solorio T, Montes-y Gómez M, González FA. Gated multimodal units for information fusion. arXiv:1702.01992. 2017. doi:10.48550/arXiv.1702.01992.
54. Cascante-Bonilla P, Sitaraman K, Luo M, Ordonez V. Moviescope: large-scale analysis of movies using multiple modalities. arXiv:1908.03180. 2019. doi:10.48550/arXiv.1908.03180.
55. Zhang ML, Zhou ZH. A review on multi-label learning algorithms. IEEE Transact Knowl Data Eng. 2013;26(8):1819–37. doi:10.1109/TKDE.2013.39.
56. Jain S, Jain SK, Vasal S. An effective TF-IDF model to improve the text classification performance. In: 2024 IEEE 13th International Conference on Communication Systems and Network Technologies (CSNT); Jabalpur, India: IEEE; 2024. p. 1–4. doi:10.1109/CSNT60213.2024.10545818.
57. Yadav A, Vishwakarma DK. A unified framework of deep networks for genre classification using movie trailer. Appl Soft Comput. 2020;96(6):106624. doi:10.1016/j.asoc.2020.106624.

58. Mervitz J, de Villiers J, Jacobs J, Kloppers M. Comparison of early and late fusion techniques for movie trailer genre labelling. In: 2020 IEEE 23rd International Conference on Information Fusion (FUSION); Online: IEEE; 2020. p. 1–8. doi:10.23919/FUSION45008.2020.9190344.
59. Hiranandani G, Vijitbenjaronk W, Koyejo S, Jain P. Optimization and analysis of the pAp@ k metric for recommender systems. In: International Conference on Machine Learning; Online; 2020. p. 4260–70. [cited 2025 Feb 1]. Available from: https://proceedings.mlr.press/v119/hiranandani20a.html
60. Makita E, Lenskiy A. A movie genre prediction based on Multivariate Bernoulli model and genre correlations. arXiv:1604.08608. 2016. doi:10.48550/arXiv.1604.08608.
61. Unal FZ, Guzel MS, Bostanci E, Acici K, Asuroglu T. Multilabel genre prediction using deep-learning frameworks. Appl Sci. 2023;13(15):8665. doi:10.3390/app13158665.