UiA University of Agder

Master

Multimodal Deep Learning for Genre Classification

Systematic Comparison of Multimodal Fusion Strategies for Multi-Label Movie Genre Classification on the MM-IMDb Dataset (Track C: Multimodal (Text + Vision))

YASIN HESSNAWI
ANWAR DEBES

SANDER RIISØEN JYHNE

University of Agder, 2025
Faculty of Engineering and Science
Department of Engineering and Sciences

Obligatorisk gruppeerklæring

Den enkelte student er selv ansvarlig for å sette seg inn i hva som er lovlige hjelpemidler, retningslinjer for bruk av disse og regler om kildebruk. Erklæringen skal bevisstgjøre studentene på deres ansvar og hvilke konsekvenser fusk kan medføre. Manglende erklæring fritar ikke studentene fra sitt ansvar.

1.	Vi erklærer herved at vår besvarelse er vårt eget arbeid, og at vi ikke har brukt andre kilder eller har mottatt annen hjelp enn det som er nevnt i besvarelsen.	Ja
2.	**Vi erklærer videre at denne besvarelsen:** • Ikke har vært brukt til annen eksamen ved annen avdeling/universitet/høgskole innenlands eller utenlands. • Ikke refererer til andres arbeid uten at det er oppgitt. • Ikke refererer til eget tidligere arbeid uten at det er oppgitt. • Har alle referansene oppgitt i litteraturlisten. • Ikke er en kopi, duplikat eller avskrift av andres arbeid eller besvarelse.	Ja
3.	Vi er kjent med at brudd på ovennevnte er å betrakte som fusk og kan medføre annullering av eksamen og utestengelse fra universiteter og høgskoler i Norge, jf. Universitets- og høgskoleloven §§4-7 og 4-8 og Forskrift om eksamen §§ 31.	Ja
4.	Vi er kjent med at alle innleverte oppgaver kan bli plagiatkontrollert.	Ja
5.	Vi er kjent med at Universitetet i Agder vil behandle alle saker hvor det forligger mistanke om fusk etter høgskolens retningslinjer for behandling av saker om fusk.	Ja
6.	Vi har satt oss inn i regler og retningslinjer i bruk av kilder og referanser på biblioteket sine nettsider.	Ja
7.	Vi har i flertall blitt enige om at innsatsen innad i gruppen er merkbart forskjellig og ønsker dermed å vurderes individuelt. Ordinært vurderes alle deltakere i prosjektet samlet.	Nei

Acknowledgements

We acknowledge the valuable computational resources provided by the University of Agder, particularly access to NVIDIA Tesla V100 GPU infrastructure, which made our experimental work feasible within the project timeline.

We express our appreciation for the availability of language assistance tools that helped refine the presentation of our technical work. The AI-powered writing assistance from the University of Oslo's ChatGPT access and Overleaf's editorial features provided useful suggestions for improving clarity and maintaining consistent academic style. We emphasize that all research concepts, experimental design, analysis, and conclusions presented in this report are entirely our own work. The language tools served only to support the readability and professional presentation of our original ideas and findings.

Finally, we thank the creators of the MM-IMDb dataset and the developers of open-source deep learning frameworks (PyTorch, Hugging Face Transformers) that enabled this research.

Abstract

Movie genre classification presents a challenging multi-label classification problem where films simultaneously belong to multiple genres. This project presents a systematic comparison of multimodal fusion strategies for movie genre classification on the MM-IMDb dataset, which combines movie plot summaries with poster images across 23 genre categories.

We implement and evaluate seven model architectures spanning three complexity levels: single-modality baselines (LSTM for text, ResNet-18 for images), transformer-based unimodal models (BERT, Vision Transformer), and three multimodal fusion approaches (Concatenation Fusion, Late Fusion, Attention Fusion). Each model is trained and evaluated using consistent experimental protocols, enabling systematic comparison of architectural trade-offs.

Our best model, Attention Fusion combining BERT and ResNet-18, achieves 59.79% macro F1-score, 65.92% micro F1-score, and 27.34% subset accuracy. This represents competitive performance with simpler architectures compared to more complex state-of-the-art approaches. The current benchmark on MM-IMDb is MM-GATBT at 64.5% macro F1, which employs Graph Attention Networks and EfficientNet. While our result falls 4.71 percentage points below this benchmark, our systematic exploration provides valuable insights into fusion strategy trade-offs.

Key findings demonstrate that encoder quality contributes substantially more to performance than fusion mechanism complexity. Upgrading from LSTM to BERT yields +24.1 percentage points improvement, while advancing from Concatenation to Attention Fusion adds only +0.6 percentage points. Additionally, simple Late Fusion achieves 99.4% of Attention Fusion's performance while being faster and more interpretable, suggesting diminishing returns from architectural complexity for this task.

Analysis of per-genre performance reveals that text modality dominates classification performance due to the direct semantic connection between plot summaries and genre labels, while visual modality provides complementary information that improves overall results by 2.78 percentage points. The models achieve strong performance on frequent genres (Drama: 76.8% F1, Comedy: 71.2% F1) but struggle with rare categories (Film-Noir: 21.7% F1, Western: 38.9% F1), reflecting the inherent class imbalance in the dataset.

This work demonstrates that systematic architectural comparison provides valuable educational insights into multimodal learning fundamentals, revealing practical guidelines for fusion strategy selection and highlighting the importance of encoder quality over fusion complexity.

Keywords: Multimodal Learning, Genre Classification, Fusion Strategies, BERT, Vision Transformer, MM-IMDb

Code Availability: The complete implementation, including all model architectures, training scripts, and evaluation code, is publicly available at:
https://github.com/yasinhessnawi1/ml-project

Contents

List of Figures

List of Acronyms

Acronym	Full Form
AI	Artificial Intelligence
AMP	Automatic Mixed Precision
AP	Average Precision
AUC	Area Under Curve
BCE	Binary Cross-Entropy
BERT	Bidirectional Encoder Representations from Transformers
CNN	Convolutional Neural Network
CUDA	Compute Unified Device Architecture
DistilBERT	Distilled BERT
F1	F1 Score (Harmonic Mean of Precision and Recall)
FN	False Negative
FP	False Positive
GloVe	Global Vectors for Word Representation
GMU	Gated Multimodal Unit
GPU	Graphics Processing Unit
HOG	Histogram of Oriented Gradients
LSTM	Long Short-Term Memory
MLP	Multi-Layer Perceptron
NLP	Natural Language Processing
pp	Percentage Points
PR	Precision-Recall
ReLU	Rectified Linear Unit
ResNet	Residual Network
RGB	Red Green Blue
ROC	Receiver Operating Characteristic
RQ	Research Question
SIFT	Scale-Invariant Feature Transform
SOTA	State-of-the-Art
SVM	Support Vector Machine
TF-IDF	Term Frequency-Inverse Document Frequency
TN	True Negative
TP	True Positive
VGG	Visual Geometry Group

Note: This list includes all acronyms used throughout the report in alphabetical order.

List of Tables

Chapter 1

Introduction

Classifying movies into genres is a fundamental task in organizing multimedia content and building recommendation systems. Traditional methods work with only one data type at a time, either textual information (plot summaries, reviews) or visual content (posters, trailers). However, movies naturally combine multiple information types: story, visual style, and themes work together to define genre. This project explores whether combining text and images improves genre classification.

1.1 Motivation

Genre classification has practical applications across the entertainment industry, from automatic content tagging on streaming platforms to recommendation systems. The challenge lies in fuzzy genre boundaries and multi-label complexity: many movies belong to multiple genres simultaneously. For example, a science fiction thriller exhibits visual elements from both genres, while plot summaries may emphasize different aspects depending on marketing strategy.

Recent deep learning advances demonstrate strong performance in both natural language processing and computer vision. Deep learning methods outperform traditional approaches like support vector machines,[1] while convolutional neural networks achieve human-level accuracy on image classification.[2] These results suggest that combining deep learning across modalities could improve genre classification.

Multimodal fusion techniques leverage complementarity between data sources.[3] While individual modalities provide partial or noisy signals, combining them produces more reliable predictions. For genre classification, plot summaries capture story structure and themes, while posters convey visual style and genre-associated symbols.

[1] Shervin Minaee et al. "Deep learning based text classification: A comprehensive review." In: *ACM Computing Surveys* 54.3 (2020), pp. 1–40. URL: https://arxiv.org/pdf/2004.03705; Qian Li et al. "A Survey on Text Classification: From Traditional to Deep Learning." In: *ACM Transactions on Intelligent Systems and Technology* 13.2 (2022), pp. 1–41. URL: https://dl.acm.org/doi/full/10.1145/3495162.

[2] Ning Zhang et al. "Comparative analysis of image classification algorithms based on traditional machine learning and deep learning." In: *Pattern Recognition Letters* 141 (2020), pp. 61–67. URL: https://www.sciencedirect.com/science/article/abs/pii/S0167865520302981; Waseem Rawat and Zenghui Wang. "Deep convolutional neural network based medical image classification for disease diagnosis." In: *Journal of Big Data* 6.1 (2019), pp. 1–18. URL: https://journalofbigdata.springeropen.com/articles/10.1186/s40537-019-0276-2.

[3] Jing Gao et al. "A survey on deep learning for multimodal data fusion." In: *Neural Computation* 32.5 (2020), pp. 829–864. URL: https://direct.mit.edu/neco/article/32/5/829/95591/A-Survey-on-Deep-Learning-for-Multimodal-Data; Dhanesh Ramachandram and Graham W Taylor. "Deep multimodal learning: A survey on recent advances and trends." In: *IEEE Signal Processing Magazine* 34.6 (2017), pp. 96–108. URL: https://ieeexplore.ieee.org/document/8103116.

1.2 Problem Statement

Despite multimodal movie datasets being available, key questions remain. First, we do not know how much text and vision each contribute to classification accuracy. Second, we must investigate optimal methods for combining multimodal information. Early fusion may struggle with heterogeneous data types, while late fusion might miss cross-modal interactions. Attention-based fusion offers a middle ground but adds complexity.

This work addresses these questions through experiments on the MM-IMDB dataset,[4] containing posters, plot summaries, and genre labels for over 25,000 films across 23 categories. The dataset presents real-world challenges: severe class imbalance (Drama: 54% vs Short: <1%) and multi-label complexity (average 2.5 genres per movie).

1.3 Research Questions

RQ1: Does multimodal fusion improve genre classification accuracy compared to using only text or only images?

RQ2: Which fusion strategy (early, late, or attention-based) gives the best performance on multi-label genre classification?

RQ3: Which modality (text or vision) is more informative for genre prediction?

RQ4: How does transfer learning from pre-trained models compare to training from scratch?

1.4 Contributions

This project provides systematic exploration of multimodal learning for movie genre classification, implementing seven model architectures that span from single-modality baselines to advanced fusion mechanisms. Our best model achieves 59.79% macro F1-score, demonstrating competitive performance while providing clear insights into architectural trade-offs.

While our result falls short of the current state-of-the-art MM-GATBT (64.5% macro F1, Seo et al., 2022[5]), which employs Graph Attention Networks and EfficientNet, our systematic comparison offers valuable educational and practical insights. The key contributions of this work are:

- **Systematic architectural comparison**: Implementation and evaluation of seven models across three complexity tiers, enabling direct comparison of fusion strategies under consistent experimental conditions.
- **Evidence-based fusion guidelines**: Demonstration that simple Late Fusion achieves 99.4% of Attention Fusion performance, suggesting diminishing returns from architectural complexity and providing practical guidance for model selection.
- **Encoder quality insights**: Clear evidence that encoder improvements (+24.1 pp from LSTM to BERT) substantially outweigh fusion mechanism refinements (+0.6 pp

[4] John Arevalo et al. "Gated Multimodal Units for Information Fusion." In: *5th International Conference on Learning Representations Workshop*. Dataset available at: http://lisi1.unal.edu.co/mmimdb/. 2017. URL: https://arxiv.org/abs/1702.01992.

[5] Seung Byum Seo, Hyoungwook Nam, and Payam Delgosha. "MM-GATBT: Enriching Multimodal Representation Using Graph Attention Network." In: *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Student Research Workshop*. Hybrid: Seattle, Washington + Online: Association for Computational Linguistics, July 2022, pp. 106–112. DOI: 10.18653/v1/2022.naacl-srw.14. URL: https://aclanthology.org/2022.naacl-srw.14/.

from Concatenation to Attention), indicating where optimization efforts should focus.

- **Modality contribution analysis**: Quantification showing text dominates performance (BERT alone: 57.01%) while vision provides meaningful complementary signal (+2.78 pp improvement), informing resource allocation decisions.
- **Per-genre performance characterization**: Detailed analysis across all 23 genres revealing performance patterns related to frequency, semantic clarity, and visual distinctiveness.

These contributions provide both theoretical understanding of multimodal fusion dynamics and practical guidelines for practitioners working with similar text-image classification tasks.

1.5 Report Organization

Chapter 2 reviews multimodal learning and genre classification. Chapter 3 introduces theoretical concepts including recurrent networks, transformers, and fusion strategies. Chapter 4 details experimental setup and model architectures. Chapter 5 presents performance comparisons. Chapter 6 interprets results and discusses implications. Chapter 7 summarizes findings and future directions.

Chapter 2

Literature Review

The field of multimodal learning has developed rapidly over the past decade, driven by advances in both deep learning architectures and the availability of large-scale datasets combining multiple data types. This chapter reviews the development of methods for genre classification, starting with traditional approaches and progressing through modern deep learning techniques to current multimodal fusion strategies.

2.1 Traditional Approaches to Classification

Before the rise of deep learning, text and image classification relied on carefully designed features combined with classical machine learning algorithms. For text analysis, methods like TF-IDF (Term Frequency-Inverse Document Frequency) represented documents as vectors based on word occurrence statistics. These vectors were then processed by classifiers such as Support Vector Machines, Naive Bayes, or Logistic Regression.[1] Similarly, image classification used hand-crafted visual features like SIFT (Scale-Invariant Feature Transform) or HOG (Histogram of Oriented Gradients), which were then fed to classifiers.[2] While these methods worked reasonably well on smaller datasets, they required significant domain expertise to design good features and struggled to scale to more complex tasks.

2.2 Deep Learning for Text Classification

The introduction of deep learning changed how text classification is approached. Instead of manually designing features, neural networks learn useful representations directly from data. Early work used recurrent neural networks, particularly Long Short-Term Memory (LSTM) networks, which could process sequences of words while maintaining information about context.[3] These models showed clear improvements over traditional methods but still faced challenges with very long sequences and training efficiency.

A major breakthrough came with the introduction of BERT (Bidirectional Encoder Representations from Transformers) by Devlin et al..[4] BERT uses a transformer architecture that processes text bidirectionally, meaning it considers both left and right context for each word

[1] Minaee et al., "Deep learning based text classification: A comprehensive review."

[2] Zhang et al., "Comparative analysis of image classification algorithms based on traditional machine learning and deep learning."

[3] Minaee et al., "Deep learning based text classification: A comprehensive review."

[4] Jacob Devlin et al. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. 2019, pp. 4171–4186. URL: https://aclanthology.org/N19-1423/.

simultaneously. The model is first pre-trained on massive amounts of unlabeled text using self-supervised tasks, then fine-tuned for specific classification problems. This approach achieved state-of-the-art results across many natural language processing benchmarks and has become a standard method for text classification tasks.

2.3 Deep Learning for Image Classification

Similar progress occurred in computer vision. Convolutional Neural Networks (CNNs) became the standard approach for image classification after achieving breakthrough results on the ImageNet challenge.[5] CNNs use layers of filters that detect patterns at different scales, starting with simple edges and building up to complex objects. The key advantage over traditional methods is that these filters are learned automatically during training rather than being hand-designed.

He et al.[6] introduced Residual Networks (ResNet), which addressed the problem of training very deep networks. Their key innovation was skip connections that allow information to flow directly across layers, making it easier to train networks with hundreds of layers. ResNet achieved significant performance improvements on image classification tasks and remains widely used today. The availability of pre-trained ResNet models on ImageNet has made transfer learning practical, where models trained on large generic datasets can be adapted to specific tasks with limited data.

2.4 Multimodal Fusion Strategies

As individual modalities improved, researchers began exploring how to effectively combine them. Gao et al.[7] provide a review of multimodal fusion methods, which are typically categorized based on when the fusion occurs. Early fusion combines raw features from different modalities before processing, late fusion combines predictions from separate unimodal models, and intermediate fusion combines learned representations at various depths in the network. Each approach has trade-offs: early fusion can capture cross-modal interactions but may struggle with heterogeneous data types, while late fusion is simpler to implement but might miss important interactions between modalities.

2.5 Genre Classification on MM-IMDb

The MM-IMDb dataset was introduced by Arevalo et al.[8] specifically for multimodal movie genre classification. Their Gated Multimodal Unit (GMU) used multiplicative gates to learn how much each modality should contribute to the final prediction. The GMU outperformed both single-modality baselines and simple fusion strategies like concatenation or averaging. Subsequent work has built on this dataset to explore different fusion mechanisms.

Li et al.[9] proposed incorporating knowledge graphs that capture relationships between movie metadata (directors, actors, release years) to improve fusion. Their model used attention

[5] Zhang et al., "Comparative analysis of image classification algorithms based on traditional machine learning and deep learning."

[6] Kaiming He et al. "Deep Residual Learning for Image Recognition." In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016, pp. 770–778. URL: https://arxiv.org/abs/1512.03385.

[7] Gao et al., "A survey on deep learning for multimodal data fusion."

[8] Arevalo et al., "Gated Multimodal Units for Information Fusion."

[9] Jiaqi Li et al. "Incorporating Domain Knowledge Graph into Multimodal Movie Genre Classification with Self-Supervised Attention and Contrastive Learning." In: *Proceedings of the 31st ACM International Conference on Multimedia*. 2023, pp. 8220–8230. URL: https://arxiv.org/abs/2310.08032.

mechanisms to weight the contribution of different modalities and achieved state-of-the-art results on MM-IMDb. Other recent work has explored graph neural networks[10] and different attention mechanisms[11] for combining text and image features. These studies consistently show that multimodal approaches outperform unimodal baselines, though the text modality typically contributes more to performance than the visual modality for this particular task.

2.6 Identified Challenges

Despite progress, several challenges remain in multimodal genre classification. First, severe class imbalance in the dataset makes it difficult for models to learn rare genres effectively. Second, the multi-label nature of the problem (movies can belong to multiple genres) requires careful design of both loss functions and evaluation metrics. Third, determining the optimal fusion strategy depends on the specific characteristics of the modalities and task, and no single approach works best in all cases. Finally, most existing work focuses on relatively simple fusion mechanisms and does not fully explore how text and visual information interact semantically for genre prediction.

This work builds on these foundations by systematically comparing different fusion strategies (early, late, and attention-based fusion) and carefully analyzing the contribution of each modality to classification performance. While we do not claim to address all these challenges, our experimental design allows us to better understand the trade-offs between different architectural choices for this task.

[10] Seung Byum Kim et al. "MM-GATBT: Enriching Multimodal Representation Using Graph Attention Network." In: *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*. 2022, pp. 105–112. URL: https://aclanthology.org/2022.naacl-srw.14.pdf.

[11] Francisco Braz, Kelwin Fernandes, and Jaime S Cardoso. "Image-Text Integration Using a Multimodal Fusion Network Module for Movie Genre Classification." In: *2021 International Conference on Content-Based Multimedia Indexing*. IEEE. 2021, pp. 1–6. URL: https://ieeexplore.ieee.org/document/9569001/.

Chapter 3

Theory and Background

This chapter presents the theoretical foundations underlying the models and methods used in this work. We begin with text representation learning approaches, including recurrent networks and transformer-based models. We then cover convolutional neural networks for image processing, followed by multimodal fusion strategies. Throughout, we focus on the specific architectures and techniques employed in our experiments on the MM-IMDb dataset.

3.1 Text Representation Learning

Text classification requires converting discrete sequences of words into continuous numerical representations that capture semantic meaning. This section covers two main approaches used in this work: recurrent neural networks with attention and transformer-based models.

3.1.1 Word Embeddings

Before processing text through neural networks, words must be converted to dense vector representations. Word embeddings map each word in a vocabulary to a fixed-dimensional vector (typically 300 dimensions). Two common approaches exist: training embeddings from scratch as part of the model, or using pre-trained embeddings like GloVe (Global Vectors for Word Representation). Pre-trained embeddings capture semantic relationships learned from large text corpora, where similar words have similar vector representations.

For a vocabulary of size V and embedding dimension d, the embedding layer is represented as a matrix $E \in \mathbb{R}^{V \times d}$. Given a sequence of word indices $[w_1, w_2, ..., w_n]$, the embedding layer produces a sequence of vectors $[e_1, e_2, ..., e_n]$ where $e_i = E[w_i]$.

3.1.2 Long Short-Term Memory Networks

Long Short-Term Memory (LSTM) networks are a type of recurrent neural network designed to capture long-range dependencies in sequential data. Unlike simple recurrent networks that struggle with vanishing gradients, LSTMs use a gating mechanism to control information flow.

An LSTM cell at time step t receives input x_t and previous hidden state h_{t-1} and cell state c_{t-1}. It computes:

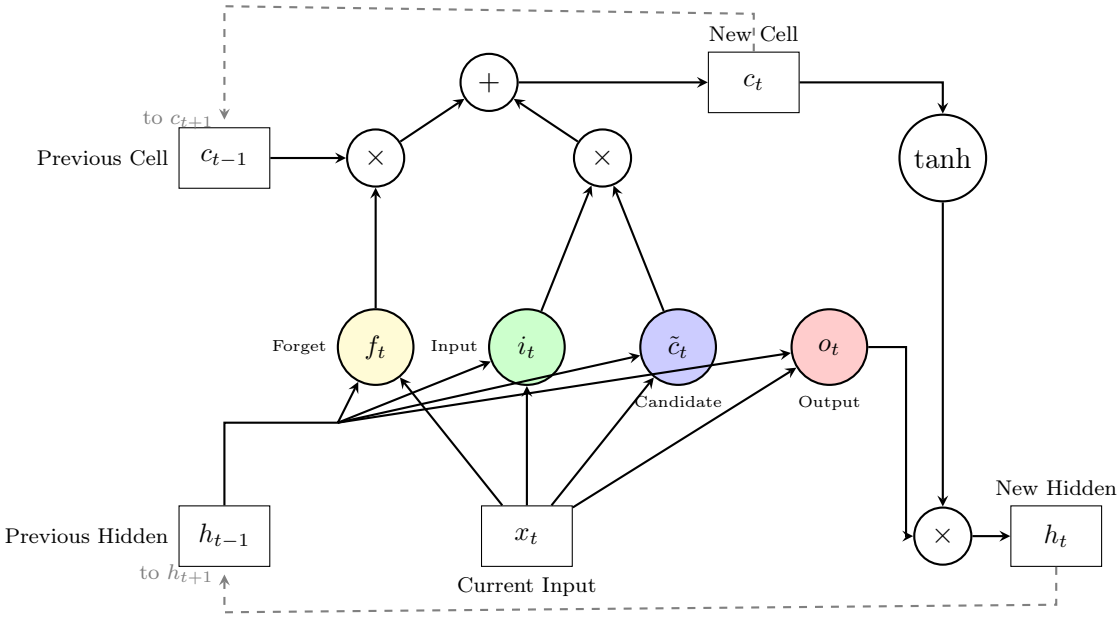

Figure 3.1: LSTM cell architecture showing the three gates (forget, input, output) and how they control information flow through the cell state c_t and hidden state h_t. The dashed arrows show how the current states feed into the next time step. The forget gate controls what to discard from the previous cell state, the input gate controls what new information to add, and the output gate determines what to output based on the updated cell state.

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \quad \text{(forget gate)} \tag{3.1}$$
$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \quad \text{(input gate)} \tag{3.2}$$
$$\tilde{c}_t = \tanh(W_c \cdot [h_{t-1}, x_t] + b_c) \quad \text{(candidate cell state)} \tag{3.3}$$
$$c_t = f_t \odot c_{t-1} + i_t \odot \tilde{c}_t \quad \text{(cell state update)} \tag{3.4}$$
$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \quad \text{(output gate)} \tag{3.5}$$
$$h_t = o_t \odot \tanh(c_t) \quad \text{(hidden state)} \tag{3.6}$$

where σ is the sigmoid function, tanh is the hyperbolic tangent, $\odot$ denotes element-wise multiplication, and W and b are learned weight matrices and bias vectors.

The forget gate f_t decides what information to discard from the previous cell state. The input gate i_t controls what new information to add. The output gate o_t determines what to output based on the cell state. This gating mechanism allows LSTMs to maintain information over long sequences.

Bidirectional LSTM: To capture context from both past and future words, we use bidirectional LSTMs that process the sequence in both forward and backward directions. The final representation at each position concatenates the forward and backward hidden states: $h_t = [h_t^{\rightarrow}, h_t^{\leftarrow}]$.

Attention Mechanism: Even with bidirectional processing, a single fixed-size vector may not capture all important information in long sequences. An attention mechanism computes a weighted sum of all hidden states, where weights indicate the importance of each position:

$$\alpha_t = \frac{\exp(e_t)}{\sum_{i=1}^{n} \exp(e_i)} \tag{3.7}$$

$$e_t = v^T \tanh(W h_t) \tag{3.8}$$

$$\text{context} = \sum_{t=1}^{n} \alpha_t h_t \tag{3.9}$$

where v and W are learned parameters, and α_t are attention weights that sum to 1.

3.1.3 Transformer-Based Models: BERT

While LSTMs process sequences step-by-step, transformers process all positions in parallel using self-attention. BERT (Bidirectional Encoder Representations from Transformers) is a pre-trained language model that has achieved strong results on many text classification tasks.

Self-Attention: The core mechanism in transformers is self-attention, which computes relationships between all pairs of words in a sequence. For an input sequence represented as matrix X, self-attention computes:

$$Q = XW^Q, \quad K = XW^K, \quad V = XW^V \tag{3.10}$$

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{3.11}$$

where W^Q, W^K, and W^V are learned projection matrices, and d_k is the dimension of the key vectors. The scaling by $\sqrt{d_k}$ prevents the dot products from becoming too large.

Intuitively, this mechanism computes how much each word should attend to every other word. The queries Q represent what each word is looking for, keys K represent what each word offers, and values V represent the actual information to retrieve.

Multi-Head Attention: BERT uses multiple attention heads that learn different relationships:

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, ..., \text{head}_h)W^O \tag{3.12}$$

where each head computes attention independently with its own parameters.

Pre-training and Fine-tuning: BERT is first pre-trained on large text corpora using two self-supervised tasks: masked language modeling (predicting randomly masked words) and next sentence prediction. This pre-training learns general language representations. For our genre classification task, we fine-tune BERT by adding a classification layer on top and training on our labeled data. We use DistilBERT, a smaller and faster version that retains 97% of BERT's performance while having 40% fewer parameters.

3.2 Image Representation Learning

Image classification requires extracting visual features that capture relevant patterns for the task. Convolutional neural networks have become the standard approach, learning hierarchical representations from raw pixels.

3.2.1 Convolutional Neural Networks

CNNs use convolution operations to detect local patterns in images. A convolution applies a filter (kernel) across the image:

$$(I * K)[i, j] = \sum_m \sum_n I[i + m, j + n] \cdot K[m, n] \tag{3.13}$$

where I is the input image, K is the kernel, and $*$ denotes convolution.

Each convolutional layer contains multiple filters that detect different patterns. Early layers learn simple patterns like edges and corners, while deeper layers combine these to detect complex objects.

Activation Functions: After each convolution, we apply a non-linear activation function. We use ReLU (Rectified Linear Unit):

$$\text{ReLU}(x) = \max(0, x) \tag{3.14}$$

ReLU is simple, computationally efficient, and helps prevent vanishing gradients during training.

Pooling: Pooling layers reduce spatial dimensions and provide translation invariance. Max pooling takes the maximum value in each region:

$$\text{MaxPool}(X)[i, j] = \max_{m,n \in \text{region}} X[i + m, j + n] \tag{3.15}$$

Batch Normalization: To stabilize training, batch normalization normalizes activations:

$$\hat{x} = \frac{x - \mu_B}{\sqrt{\sigma_B^2 + \epsilon}} \tag{3.16}$$

where μ_B and σ_B are the mean and standard deviation of the current batch, and ϵ is a small constant for numerical stability.

3.2.2 Residual Networks (ResNet)

As networks become deeper, they can suffer from degradation where training error increases with depth. ResNet addresses this with skip connections that allow information to bypass layers.

A residual block learns the residual function $F(x)$ instead of the direct mapping $H(x)$:

$$H(x) = F(x) + x \tag{3.17}$$

where x is the input and $F(x)$ represents the stacked convolutional layers. The skip connection adds x directly to the output. This makes it easier for the network to learn identity mappings when needed, as it only needs to set $F(x) = 0$.

Transfer Learning with ResNet: We use ResNet-18 pre-trained on ImageNet, which contains 1.2 million images across 1000 categories. The early layers have learned general visual features like edges and textures. For our task, we replace the final classification layer and fine-tune the network on movie posters. We can choose to freeze early layers and only train deeper layers, or fine-tune the entire network with a small learning rate.

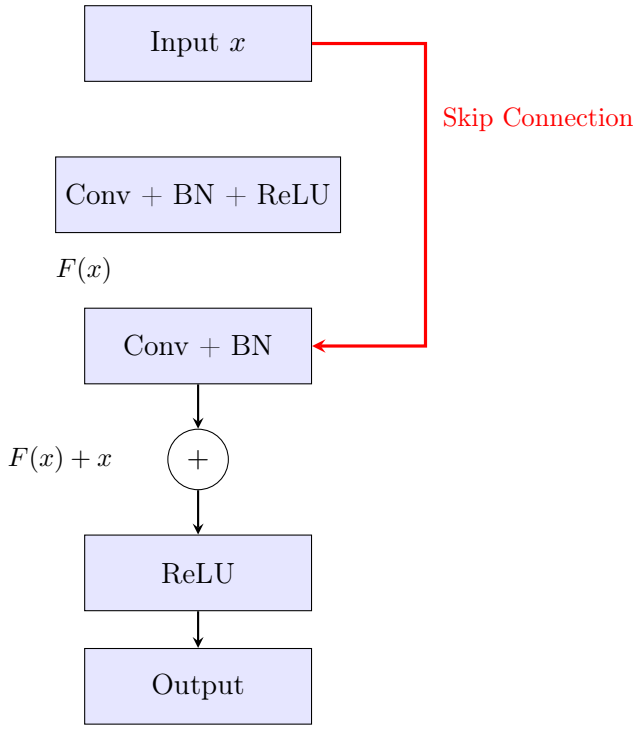

Figure 3.2: Residual block in ResNet. The skip connection (red arrow) allows the input x to bypass the convolutional layers, making it easier to learn identity mappings and train very deep networks.

3.3 Multimodal Fusion Strategies

Combining information from text and images requires careful design of fusion mechanisms. Different strategies make different assumptions about how modalities interact.

3.3.1 Early Fusion

Early fusion combines features from different modalities before the final classification. After encoding text and images separately, we project them to a common dimension and concatenate:

$$h_{\text{text}} = f_{\text{text}}(x_{\text{text}}) \in \mathbb{R}^{d_t} \tag{3.18}$$

$$h_{\text{vision}} = f_{\text{vision}}(x_{\text{vision}}) \in \mathbb{R}^{d_v} \tag{3.19}$$

$$h_{\text{fused}} = [W_t h_{\text{text}}; W_v h_{\text{vision}}] \in \mathbb{R}^{d} \tag{3.20}$$

where f_{text} and f_{vision} are the text and vision encoders, W_t and W_v are projection matrices, and $[;]$ denotes concatenation. The fused representation then passes through additional layers:

$$y = \text{sigmoid}(W_3 \cdot \text{ReLU}(W_2 \cdot \text{ReLU}(W_1 h_{\text{fused}}))) \tag{3.21}$$

where W_1, W_2, W_3 are weight matrices with dropout applied between layers.

Advantages: Early fusion allows the model to learn joint representations and cross-modal interactions.

Challenges: Different modalities may have very different distributions, making it hard to learn a good joint representation. The model must learn to balance their contributions.

3.3.2 Late Fusion

Late fusion processes each modality independently and combines their predictions:

$$p_{\text{text}} = f_{\text{text}}(x_{\text{text}}) \in [0,1]^C \tag{3.22}$$
$$p_{\text{vision}} = f_{\text{vision}}(x_{\text{vision}}) \in [0,1]^C \tag{3.23}$$
$$p_{\text{final}} = \alpha \cdot p_{\text{text}} + (1-\alpha) \cdot p_{\text{vision}} \tag{3.24}$$

where C is the number of classes (genres) and α controls the weight of each modality. We can set α as a fixed hyperparameter or learn it during training.

Advantages: Simple to implement and each modality is processed independently, making it easier to understand their individual contributions.

Challenges: Cannot learn cross-modal interactions. If text and vision provide complementary information, late fusion may miss these interactions.

3.3.3 Attention-Based Fusion

Attention fusion uses cross-attention to let text and vision features attend to each other:

$$Q_t = h_{\text{text}} W^Q, \quad K_v = h_{\text{vision}} W^K, \quad V_v = h_{\text{vision}} W^V \tag{3.25}$$
$$\text{Attended}_t = \text{softmax}\left(\frac{Q_t K_v^T}{\sqrt{d_k}}\right) V_v \tag{3.26}$$

Similarly, vision attends to text. The attended representations are then combined and passed to a classifier.

Advantages: Can learn which parts of one modality are relevant to parts of another. Provides interpretability through attention weights.

Challenges: More parameters to learn and increased computational cost.

3.4 Training Deep Neural Networks

Training multimodal models requires careful choice of optimization algorithms, loss functions, and regularization techniques.

3.4.1 Optimization: AdamW

We use AdamW (Adam with decoupled weight decay), which combines adaptive learning rates with proper L2 regularization. For each parameter θ, AdamW maintains exponential moving averages of gradients and squared gradients:

$$m_t = \beta_1 m_{t-1} + (1 - \beta_1) g_t \tag{3.27}$$

$$v_t = \beta_2 v_{t-1} + (1 - \beta_2) g_t^2 \tag{3.28}$$

$$\hat{m}_t = \frac{m_t}{1 - \beta_1^t}, \quad \hat{v}_t = \frac{v_t}{1 - \beta_2^t} \tag{3.29}$$

$$\theta_t = \theta_{t-1} - \eta \left(\frac{\hat{m}_t}{\sqrt{\hat{v}_t} + \epsilon} + \lambda \theta_{t-1} \right) \tag{3.30}$$

where g_t is the gradient, η is the learning rate, $\beta_1 = 0.9$ and $\beta_2 = 0.999$ are decay rates, $\epsilon = 10^{-8}$ prevents division by zero, and λ is the weight decay coefficient. The key difference from standard Adam is that weight decay is applied directly to parameters rather than included in the gradient.

3.4.2 Loss Function for Multi-Label Classification

Movie genre classification is a multi-label problem where each movie can belong to multiple genres. We use Binary Cross-Entropy (BCE) loss, which treats each genre as an independent binary classification:

$$\mathcal{L}_{\text{BCE}} = -\frac{1}{N} \frac{1}{C} \sum_{i=1}^{N} \sum_{j=1}^{C} \left[y_{ij} \log(\hat{y}_{ij}) + (1 - y_{ij}) \log(1 - \hat{y}_{ij}) \right] \tag{3.31}$$

where N is the number of samples, C is the number of genres, $y_{ij} \in \{0, 1\}$ is the true label, and $\hat{y}_{ij} = \sigma(z_{ij})$ is the predicted probability using the sigmoid function $\sigma(z) = \frac{1}{1+e^{-z}}$.

Weighted BCE: To handle class imbalance, we weight the loss for each genre by the inverse of its frequency:

$$\mathcal{L}_{\text{weighted}} = -\frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{C} w_j \left[y_{ij} \log(\hat{y}_{ij}) + (1 - y_{ij}) \log(1 - \hat{y}_{ij}) \right] \tag{3.32}$$

where $w_j = \frac{N_{\text{neg},j}}{N_{\text{pos},j}}$ gives higher weight to rare genres.

3.4.3 Regularization

Dropout: Randomly sets a fraction of activations to zero during training, preventing the model from relying too heavily on specific features. We use dropout rates between 0.3 and 0.5 in different layers.

Weight Decay: L2 regularization penalizes large weights, encouraging the model to learn simpler patterns. We use weight decay of 0.01 to 0.05.

Early Stopping: Monitors validation performance and stops training when it stops improving. We use patience of 10-20 epochs, meaning training stops if validation F1-score does not improve for that many epochs.

3.4.4 Learning Rate Scheduling

We use ReduceLROnPlateau scheduling, which reduces the learning rate when validation performance plateaus:

$$\eta_{\text{new}} = \eta_{\text{old}} \times \text{factor} \tag{3.33}$$

when no improvement is seen for a specified number of epochs (patience). This allows the model to make large updates early in training and fine-grained updates later. We use factor = 0.5 and patience = 5-7 epochs.

3.5 Summary

This chapter has covered the key theoretical concepts underlying our multimodal genre classification system. Text encoding uses either bidirectional LSTMs with attention or transformer-based BERT models. Image encoding uses convolutional neural networks, specifically ResNet with skip connections that enable training of very deep networks. Fusion strategies range from simple concatenation (early fusion) to weighted averaging (late fusion) to learned cross-attention (attention fusion). Training uses the AdamW optimizer with Binary Cross-Entropy loss adapted for multi-label classification and class imbalance. These components form the foundation for the experimental methodology described in the next chapter.

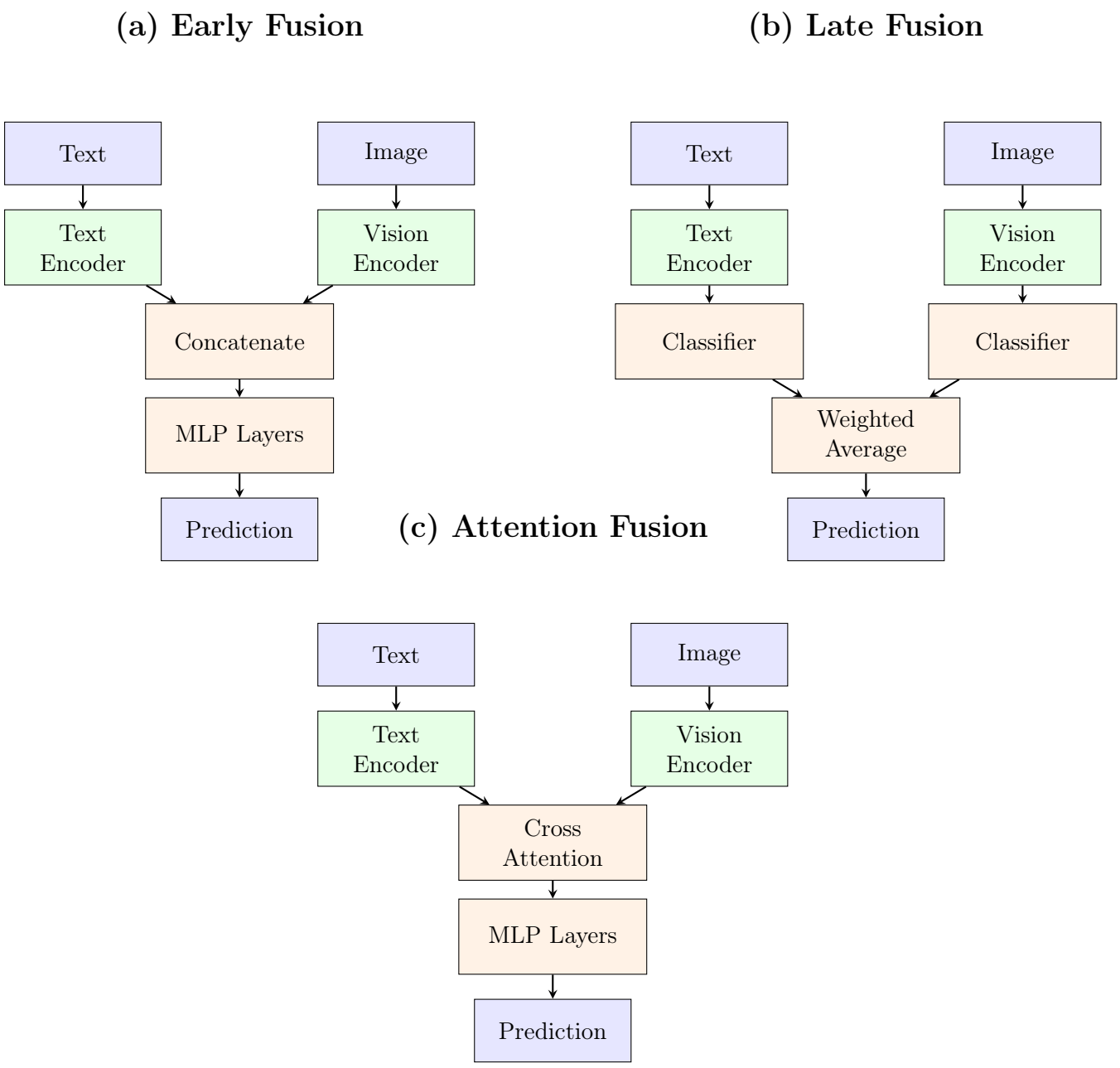

Figure 3.3: Comparison of three fusion strategies: (a) Early fusion concatenates encoded features before classification, (b) Late fusion combines independent predictions, and (c) Attention fusion uses cross-attention to learn interactions between modalities.

Chapter 4

Methods

This chapter describes the experimental methodology used to investigate multimodal genre classification on the MM-IMDb dataset. We begin with a detailed description of the dataset and preprocessing steps, followed by the architectures of all seven models implemented. We then outline the training procedure, evaluation metrics, and experimental design that addresses our research questions.

4.1 Dataset

4.1.1 MM-IMDb Dataset Overview

We use the MM-IMDb (Multimodal IMDb) dataset introduced by Arevalo et al.,[1] which contains movie metadata combining visual and textual information. The dataset was downloaded from Kaggle[2] and consists of 25,959 movies. Each movie is annotated with a movie poster (RGB image of size 256×160 pixels), a plot summary (textual description of the movie plot), genre labels (multi-label annotations from 23 possible genres), and metadata including title, year, director, and actors.

After filtering out samples with plot summaries shorter than 10 tokens, the final dataset contains 25,815 movies. We use the standard train/validation/test split of 70%/15%/15%, resulting in 18,070 training samples, 3,872 validation samples, and 3,873 test samples.

4.1.2 Genre Distribution and Class Imbalance

The dataset exhibits severe class imbalance, a common challenge in real-world multi-label classification. Table 4.1 shows the distribution of the 23 genres. Drama is the most frequent genre, appearing in 53.9% of movies, while rare genres like Short, Film-Noir, and Sport appear in less than 1% of samples. This represents an imbalance ratio of approximately 278:1 between the most and least frequent genres.

The multi-label nature adds additional complexity: each movie belongs to an average of 2.5 genres, with some movies assigned up to 8 genres. Common genre combinations include Drama+Romance (found in approximately 4,000 movies), Drama+Thriller (3,500 movies), and Comedy+Romance (2,800 movies).

[1] Arevalo et al., "Gated Multimodal Units for Information Fusion."
[2] https://www.kaggle.com/datasets/johnarevalo/mmimdb

Table 4.1: Genre distribution in the MM-IMDb dataset showing severe class imbalance.

Genre	Count	%	Genre	Count	%
Drama	13,917	53.9	Mystery	1,800	7.0
Comedy	8,549	33.1	Fantasy	1,500	5.8
Romance	5,348	20.7	Family	1,400	5.4
Thriller	5,181	20.1	Biography	1,200	4.6
Action	3,829	14.8	War	900	3.5
Horror	3,541	13.7	History	800	3.1
Documentary	2,707	10.5	Music	700	2.7
Crime	2,697	10.4	Animation	600	2.3
Sci-Fi	2,050	7.9	Musical	400	1.5
Adventure	2,032	7.9	Western	300	1.2
			Sport	632	2.4
			Film-Noir	100	0.4
			Short	50	0.2

4.1.3 Dataset Challenges

This dataset presents several challenges that make it suitable for evaluating multimodal learning approaches. The severe class imbalance requires models to avoid bias toward frequent genres while still learning rare categories effectively. Unlike single-label classification, predicting exact genre combinations is substantially harder, as evidenced by low subset accuracy scores in prior work. Plot summaries emphasize narrative and thematic content, while posters convey visual style and iconography, requiring models to effectively integrate different information types. Finally, genre boundaries are often fuzzy, and annotator disagreement can occur for movies that blend multiple genres.

4.2 Data Preprocessing

4.2.1 Text Preprocessing

Plot summaries are provided as pre-tokenized sequences in the original dataset. We apply different preprocessing pipelines depending on the model architecture. Table 4.2 summarizes the text preprocessing configurations for LSTM and BERT models.

Table 4.2: Text preprocessing configurations for LSTM and BERT models.

Configuration	LSTM Models	BERT Models
Tokenizer	Whitespace splitting	DistilBERT WordPiece
Vocabulary size	69,980 tokens	30,522 tokens
Maximum length	128 tokens (75% coverage)	512 tokens (99% coverage)
Special tokens	None	[CLS], [SEP]
Embedding initialization	GloVe 300d (6B tokens)	Pretrained DistilBERT
Padding	Applied to shorter sequences	Applied with attention masks

For LSTM models, we use the provided vocabulary of 69,980 unique tokens and truncate sequences to a maximum length of 128 tokens, which covers 75% of the data fully based on the median length of 107 tokens. We apply padding to shorter sequences and initialize word embeddings with 300-dimensional GloVe vectors[3] pre-trained on 6 billion tokens.

[3]Jeffrey Pennington, Richard Socher, and Christopher D Manning. "GloVe: Global Vectors for Word Representation." In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP).*

For BERT models, we use the DistilBERT tokenizer with WordPiece vocabulary of 30,522 tokens. We convert pre-tokenized sequences back to text, then re-tokenize using the BERT tokenizer. The maximum sequence length is set to 512 tokens (BERT's maximum), which covers 99% of the data. Special tokens [CLS] and [SEP] are added at the start and end of sequences, and we apply padding with attention masks for batch processing.

4.2.2 Image Preprocessing

Movie posters are provided as 256×160 RGB images. All images are resized to 224×224 pixels to match the standard input size for ImageNet-pretrained models. We normalize pixel values using ImageNet statistics (mean = [0.485, 0.456, 0.406], std = [0.229, 0.224, 0.225]) and convert from HWC (Height×Width×Channels) to CHW format for PyTorch.

During training, we apply data augmentation to improve model generalization. Images are randomly flipped horizontally with probability 0.5, randomly rotated up to ±5 degrees, and subjected to color jittering with brightness, contrast, and saturation adjusted by up to ±15% and hue by up to ±5%. Validation and test sets use only basic preprocessing without augmentation to ensure consistent evaluation.

4.3 Model Architectures

We implement seven models to systematically investigate the contribution of each modality and different fusion strategies. Table 4.3 provides an overview of all models.

Table 4.3: Overview of all seven models implemented in this work.

Model	**Modality**	**Type**	**Purpose**
LSTM Text	Text only	Recurrent	Baseline text model
BERT Text	Text only	Transformer	Pretrained text model
Custom CNN	Vision only	Convolutional	Baseline vision model
ResNet-18	Vision only	Residual	Pretrained vision model
Early Fusion	Text + Vision	Concatenation	Feature-level fusion
Late Fusion	Text + Vision	Weighted average	Decision-level fusion
Attention Fusion	Text + Vision	Cross-attention	Learned fusion

4.3.1 Text-Only Models

LSTM Text Model: The LSTM text model uses a bidirectional LSTM with an attention mechanism to process plot summaries. The architecture consists of a 300-dimensional embedding layer initialized with GloVe vectors, followed by a 2-layer bidirectional LSTM with 256 hidden units per direction and dropout rate of 0.3. An attention mechanism learns importance weights over all time steps, allowing the model to focus on relevant parts of the plot summary. The attended representation is projected to 23 dimensions through a linear classification head with sigmoid activation for multi-label prediction.

The model processes sequences as follows: word embeddings are fed to the bidirectional LSTM, producing forward and backward hidden states that are concatenated at each time step. The attention mechanism computes a weighted sum of these hidden states, and the resulting context vector is passed to the sigmoid-activated classifier.

2014, pp. 1532–1543. URL: https://aclanthology.org/D14-1162/.

DistilBERT Text Model: We use the pretrained DistilBERT-base-uncased model,[4] which is a distilled version of BERT retaining 97% of its performance with 40% fewer parameters. DistilBERT has 6 transformer layers, 768 hidden dimensions, and 12 attention heads, totaling approximately 66 million parameters. We fine-tune all layers with a small learning rate (3e-5) rather than freezing the base model. The [CLS] token representation from the final layer is extracted, passed through a dropout layer (rate 0.3), and fed to a linear classification head (768 $\rightarrow$ 23 dimensions) with sigmoid activation.

4.3.2 Vision-Only Models

Custom CNN: The custom CNN is a four-layer convolutional network trained from scratch. The architecture uses progressively increasing channel dimensions (64 $\rightarrow$ 128 $\rightarrow$ 256 $\rightarrow$ 512) with kernel sizes of 7×7 for the first layer and 3×3 for subsequent layers. Each convolutional layer is followed by batch normalization, ReLU activation, and 2×2 max pooling, progressively reducing spatial dimensions from 224×224 to 7×7. Global average pooling aggregates the final feature maps into a 512-dimensional vector, which is passed through dropout (rate 0.7) before the final linear classifier. The high dropout rate is crucial for preventing overfitting when training from scratch with limited data.

ResNet-18: We use ImageNet-pretrained ResNet-18[5] with approximately 11.3 million parameters. ResNet-18 employs residual connections that enable training deeper networks by mitigating the vanishing gradient problem. We adopt a partial fine-tuning strategy: early layers (layer1 and layer2) that capture low-level visual features like edges and textures are frozen, while deeper layers (layer3 and layer4) that capture high-level semantic features are fine-tuned for the genre classification task. The final fully connected layer is replaced with a linear layer (512 $\rightarrow$ 23 dimensions) with sigmoid activation, and dropout (rate 0.3) is applied before the classifier.

4.3.3 Multimodal Fusion Models

All fusion models combine DistilBERT for text and ResNet-18 for vision, as these were the best-performing unimodal models. The three fusion strategies represent different points in the spectrum from feature-level to decision-level integration.

Early Fusion concatenates text and vision features at the feature level before classification. DistilBERT produces 768-dimensional text representations while ResNet-18 produces 512-dimensional vision representations. Both features are projected to a common 512-dimensional space through linear layers, then concatenated to form a 1024-dimensional joint representation. This fused representation is passed through a multi-layer perceptron with hidden dimensions [1024, 512, 256], ReLU activations, and dropout rates [0.3, 0.3]. The final layer projects to 23 dimensions with sigmoid activation.

Late Fusion trains separate classifiers for each modality and combines their predictions through learned weighted averaging. The text branch uses DistilBERT followed by a classifier to produce 23-dimensional predictions, while the vision branch uses ResNet-18 followed by its own classifier. The final prediction is computed as $p_{\text{final}} = \alpha \cdot p_{\text{text}} + (1 - \alpha) \cdot p_{\text{vision}}$, where the weight α is a trainable parameter initialized to 0.5. Both branches are trained jointly but maintain separate decision pathways.

Attention Fusion uses cross-attention mechanisms to model interactions between modalities. DistilBERT extracts 768-dimensional text features while ResNet-18 extracts 512-

[4] Victor Sanh et al. "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter." In: *arXiv preprint arXiv:1910.01108* (2019). URL: https://arxiv.org/abs/1910.01108.

[5] He et al., "Deep Residual Learning for Image Recognition."

dimensional vision features. An 8-head cross-attention module computes attention weights that measure the relevance of each modality to the other. Text features serve as queries while vision features serve as keys and values, and vice versa in a symmetric operation. The attended representations are concatenated to form a 1024-dimensional joint representation, which is passed through a multi-layer perceptron with hidden dimension 512, ReLU activation, and dropout rate 0.1. The final layer projects to 23 dimensions with sigmoid activation.

Table 4.4 summarizes the architecture details and parameter counts for all models.

Table 4.4: Detailed architecture specifications and parameter counts.

Model	**Backbone Params**	**Head Params**	**Total Params**
LSTM Text	20.9M (embeddings)	3.1M	24.0M
BERT Text	66.4M (DistilBERT)	0.2M	66.6M
Custom CNN	–	1.7M	1.7M
ResNet-18	11.2M (ImageNet)	0.1M	11.3M
Early Fusion	77.6M (BERT+ResNet)	2.6M	80.2M
Late Fusion	77.6M (BERT+ResNet)	0.3M	77.9M
Attention Fusion	77.6M (BERT+ResNet)	6.2M	83.8M

4.4 Training Procedure

4.4.1 Optimization and Regularization

All models are trained using the AdamW optimizer.[6] Table 4.5 summarizes the complete training configuration.

Table 4.5: Training configuration for all models.

Parameter	**Value**
Optimizer	AdamW
Learning rate (from scratch)	3e-4 (LSTM, Custom CNN)
Learning rate (fine-tuning)	3e-5 (BERT, ResNet, Fusion models)
Weight decay	0.05
Beta parameters	$\beta_1 = 0.9$, $\beta_2 = 0.999$
Epsilon	10^{-8}
Batch size	32
Maximum epochs	50
Gradient clipping	Maximum norm of 1.0
Mixed precision	Enabled (AMP)
Learning Rate Scheduler (ReduceLROnPlateau)	
Monitored metric	Validation F1-macro
Patience	7 epochs
Reduction factor	0.5
Minimum learning rate	10^{-6}
Early Stopping	
Monitored metric	Validation F1-macro
Patience	10-20 epochs
Mode	Maximize

Models trained from scratch (LSTM and Custom CNN) use a higher learning rate of 3e-4, while models with pretrained components (BERT, ResNet, and all fusion models) use a lower

[6] Ilya Loshchilov and Frank Hutter. "Decoupled Weight Decay Regularization." In: *International Conference on Learning Representations*. 2019. URL: https://arxiv.org/abs/1711.05101.

learning rate of 3e-5 for careful fine-tuning. We use ReduceLROnPlateau scheduling that monitors validation F1-macro score and reduces the learning rate by a factor of 0.5 when performance does not improve for 7 consecutive epochs, with a minimum learning rate of 10^{-6}.

Regularization is applied through multiple mechanisms. Dropout is used in all models with rates ranging from 0.3 to 0.7 depending on the architecture (higher rates for models trained from scratch). Weight decay in the AdamW optimizer provides L2 regularization. Early stopping with patience of 10 to 20 epochs monitors validation F1-macro and terminates training if performance stops improving. We save the best model checkpoint based on validation performance.

4.4.2 Loss Function

For multi-label classification with class imbalance, we use weighted Binary Cross-Entropy loss (see Chapter 3) where class weights $w_j = \frac{N_{\text{neg},j}}{N_{\text{pos},j}}$ are computed from training data to give higher importance to rare genres.

4.4.3 Hardware and Implementation

All experiments were conducted on a University of Agder provided server equipped with an NVIDIA Tesla V100-SXM3-32GB GPU (Compute Capability 7.0, 31.73 GB total memory) with CUDA enabled. Table 4.6 summarizes the implementation details.

Table 4.6: Implementation details and software versions.

Component	**Version/Configuration**
Framework	PyTorch 1.13
Transformers library	Hugging Face Transformers 4.25
Random seed	42
Deterministic mode	Enabled
Mixed precision	AMP (Automatic Mixed Precision)
GPU	NVIDIA Tesla V100-SXM3-32GB
CUDA memory	31.73 GB

We set random seed 42 and enable deterministic mode in PyTorch for reproducibility. Mixed precision training using Automatic Mixed Precision (AMP) is enabled for faster training and reduced memory usage.

4.5 Evaluation Metrics

We evaluate all models using multiple metrics appropriate for multi-label classification. Our primary metric is F1-macro score, which treats all genres equally regardless of their frequency in the dataset.

4.5.1 Primary and Additional Metrics

F1-macro computes F1 score independently for each genre then averages across all genres, giving equal weight to both frequent and rare genres. This is crucial for our imbalanced dataset where macro-averaging prevents dominant genres from overshadowing rare ones. Table 4.7 summarizes all evaluation metrics used in this work. The detailed mathematical formulation is provided in Appendix A.

Table 4.7: Evaluation metrics used for multi-label classification.

Metric	Description
F1-Macro	Average F1 across all genres (equal weight) *primary metric*
F1-Micro	Aggregates predictions across all genres (frequency-weighted)
F1-Weighted	Weights each genre's F1 by its support
Precision-Macro	Average precision across all genres
Recall-Macro	Average recall across all genres
ROC-AUC-Macro	Area under ROC curve, averaged across genres
Hamming Loss	Fraction of incorrect labels
Subset Accuracy	Percentage of exact matches (all genres correct)

4.5.2 Threshold Selection

For multi-label prediction, we must convert continuous probability outputs to binary decisions using a threshold. We employ different threshold strategies for unimodal and multimodal models to ensure fair comparison.

For unimodal models (LSTM, BERT, Custom CNN, ResNet-18), we perform automated threshold optimization on the validation set. A Python script evaluates thresholds from 0.1 to 0.9 in steps of 0.02 and selects the threshold that maximizes validation F1-Macro. This procedure yielded optimal thresholds of 0.34 for LSTM Text and 0.28 for BERT Text.

For multimodal fusion models (Early, Late, Attention Fusion), we use a fixed threshold of 0.5 for all three strategies to ensure equal treatment and enable fair comparison. While this may not be optimal for each individual model, it removes threshold selection as a confounding variable when comparing fusion approaches. Manual testing confirmed that 0.5 provides balanced performance across all fusion models.

At test time, per-class F1 scores are computed to analyze performance on individual genres using the selected thresholds.

4.6 Experimental Design

Our experimental design systematically addresses the four research questions posed in Chapter 1. The seven models were chosen to enable controlled comparisons across modalities and fusion strategies.

4.6.1 Model Selection Rationale

The unimodal baselines establish individual modality performance. For text, we compare LSTM (recurrent architecture) against BERT (transformer architecture with pretraining) to evaluate the benefit of transfer learning. For vision, we compare Custom CNN (trained from scratch) against ResNet-18 (pretrained on ImageNet) to similarly assess transfer learning effectiveness.

The multimodal models test different fusion strategies while controlling for encoder architecture. Early Fusion, Late Fusion, and Attention Fusion represent different points in the spectrum from feature-level to decision-level integration. All three use identical encoders (DistilBERT for text and ResNet-18 for vision) to isolate the effect of the fusion strategy itself.

4.6.2 Mapping to Research Questions

Table 4.8 shows how our experimental design addresses each research question.

Table 4.8: Mapping of model comparisons to research questions.

Research Question	**Comparison Strategy**
RQ1: Does multimodal fusion improve performance?	Compare best multimodal model against best unimodal models
RQ2: Which fusion strategy works best?	Direct comparison of Early, Late, and Attention Fusion on same test set; analyze trade-offs in parameter count and performance
RQ3: Which modality is more informative?	Compare BERT Text (best text) vs ResNet-18 (best vision); examine per-genre performance to identify modality-specific strengths
RQ4: How does transfer learning compare to training from scratch?	Text: BERT (pretrained on BooksCorpus + Wikipedia) vs LSTM (GloVe initialization); Vision: ResNet-18 (ImageNet pretrained) vs Custom CNN (from scratch)

4.6.3 Ablation Studies

Within each model family, we validate architectural choices through ablation studies. For LSTM, we test configurations with and without the attention mechanism, and with and without GloVe initialization. For ResNet-18, we explore different fine-tuning strategies including freezing all layers versus partial fine-tuning of deeper layers. For fusion models, we experiment with different fusion dimensionalities and dropout rates to identify optimal configurations.

4.7 Summary

This chapter has described the complete experimental methodology for multimodal genre classification on MM-IMDb. The dataset presents realistic challenges including severe class imbalance (278:1 ratio) and multi-label complexity (2.5 genres per movie on average). Text is preprocessed using either LSTM-specific tokenization with GloVe embeddings or BERT WordPiece tokenization, while images undergo standard ImageNet normalization and conservative augmentation. Seven models spanning unimodal baselines, pretrained models, and three fusion strategies enable systematic investigation of our research questions. Training uses AdamW optimization with weighted BCE loss to handle class imbalance, and evaluation focuses on F1-macro as the primary metric to ensure fair treatment of rare genres. All experiments run on NVIDIA Tesla V100 GPU with PyTorch, ensuring reproducibility through fixed random seeds and deterministic operations. The next chapter presents the results of these experiments.

Chapter 5

Results

This chapter presents the experimental results from evaluating all seven models on the MM-IMDb test set. We begin with an overview of overall performance, then systematically address each research question through controlled comparisons. The analysis examines performance differences across modalities, fusion strategies, and training approaches, followed by detailed per-genre analysis and model behavior characterization.

5.1 Overall Performance

Table 5.1 presents the complete test set performance for all seven models, ranked by F1-Macro score (our primary evaluation metric). All models were evaluated on the same 3,873 test samples using the threshold selection strategies described in Section 4.5.1.

Table 5.1: Overall test set performance for all models, ranked by F1-Macro score.

Model	F1-Macro	ROC-AUC	Precision	Recall	Subset Acc.	Threshold
Attention Fusion	**59.79%**	**90.61%**	58.29%	**61.88%**	17.97%	0.50
Late Fusion	59.43%	89.78%	**59.15%**	60.35%	18.18%	0.50
Early Fusion	58.47%	89.00%	56.17%	61.73%	16.99%	0.50
BERT Text	57.01%	88.38%	58.38%	56.88%	**18.46%**	0.28
LSTM Text	43.05%	82.70%	41.74%	44.74%	9.40%	0.34
ResNet-18	29.73%	73.29%	22.15%	52.00%	1.29%	0.50
Custom CNN	24.17%	68.14%	17.00%	59.54%	0.03%	0.50

The results show a clear performance hierarchy. Attention Fusion achieves the highest F1-Macro score at 59.79%, followed closely by Late Fusion at 59.43% and Early Fusion at 58.47%. All three multimodal fusion approaches outperform the best unimodal model (BERT Text at 57.01%). Among unimodal models, text-based approaches substantially outperform vision-based models, with BERT achieving nearly double the F1-Macro of ResNet-18 (57.01% vs 29.73%). Transfer learning proves critical for both modalities, as BERT outperforms LSTM by 13.96 percentage points and ResNet-18 outperforms Custom CNN by 5.56 percentage points.

The ROC-AUC scores follow a similar pattern, with Attention Fusion achieving 90.61%, indicating strong ranking ability across all genres. Notably, while vision models show very low subset accuracy (0.03% and 1.29%), they maintain moderate recall (52.00% and 59.54%), suggesting they tend to over-predict genres rather than miss them entirely.

5.2 Research Question 1: Multimodal vs Unimodal Performance

Does multimodal fusion improve genre classification accuracy compared to using only text or only images?

Table 5.2 compares the best multimodal model against the best models from each unimodal category.

Table 5.2: Comparison of multimodal fusion against best unimodal models.

Category	Model	F1-Macro	F1-Micro	ROC-AUC	Hamming Loss
Multimodal	Attention Fusion	**59.79%**	**65.90%**	**90.61%**	**0.0770**
Best Unimodal	BERT Text	57.01%	64.74%	88.38%	0.0781
Best Text	BERT Text	57.01%	64.74%	88.38%	0.0781
Best Vision	ResNet-18	29.73%	38.43%	73.29%	0.2112
Improvements (Multimodal vs Best Unimodal)					
Absolute Gain	–	+2.78 pp	+1.16 pp	+2.23 pp	-0.0011
Relative Gain	–	+4.9%	+1.8%	+2.5%	-1.4%

The answer to RQ1 is definitively **yes**: multimodal fusion provides measurable and consistent improvement over unimodal approaches. Attention Fusion achieves 59.79% F1-Macro, representing an absolute improvement of 2.78 percentage points (4.9% relative improvement) over the best unimodal model (BERT Text). This improvement is consistent across all evaluation metrics, including F1-Micro (+1.16 pp), ROC-AUC (+2.23 pp), and Hamming Loss (-0.0011).

Importantly, all three fusion strategies outperform the best unimodal baseline. Early Fusion achieves 58.47% (+1.46 pp over BERT), Late Fusion achieves 59.43% (+2.42 pp), and Attention Fusion achieves 59.79% (+2.78 pp). The consistency of improvement across fusion methods demonstrates that the gain comes from multimodal integration itself rather than a specific fusion architecture.

The improvement is particularly notable given that the text modality already performs strongly (57.01%), while the vision modality contributes relatively weakly (29.73%). This suggests that even when one modality dominates, the weaker modality provides complementary information that refines predictions. The vision modality appears to offer additional signals about visual style, iconography, and genre-specific aesthetics that are not fully captured in plot summaries.

5.3 Research Question 2: Fusion Strategy Comparison

Which fusion strategy (early fusion, late fusion, or attention-based fusion) gives the best performance on multi-label genre classification?

Table 5.3 directly compares the three fusion strategies, all using identical encoders (DistilBERT for text and ResNet-18 for vision).

The answer to RQ2 is **Attention Fusion**, which achieves the highest F1-Macro (59.79%) and ROC-AUC (90.61%). However, the performance differences between fusion strategies are relatively small compared to the gap between multimodal and unimodal approaches. Attention Fusion outperforms Late Fusion by only 0.36 percentage points and Early Fusion by 1.32 percentage points.

Attention Fusion demonstrates the strongest overall performance through its cross-attention mechanism that dynamically weights the contribution of each modality based on input con-

Table 5.3: Comparison of three fusion strategies with identical encoders.

Fusion Strategy	F1-Macro	Precision	Recall	ROC-AUC	Parameters	Training Time
Attention Fusion	**59.79%**	58.29%	**61.88%**	**90.61%**	83.8M	1.5x
Late Fusion	59.43%	**59.15%**	60.35%	89.78%	77.9M	1.0x
Early Fusion	58.47%	56.17%	61.73%	89.00%	80.2M	1.0x
Performance Differences						
Attention vs Late	+0.36 pp	-0.86 pp	+1.53 pp	+0.83 pp	+5.9M	+50%
Attention vs Early	+1.32 pp	+2.12 pp	+0.15 pp	+1.61 pp	+3.6M	+50%
Late vs Early	+0.96 pp	+2.98 pp	-1.38 pp	+0.78 pp	-2.3M	–

tent. The 8-head attention module allows the model to learn complex interactions between text and vision features, capturing how plot summaries and visual elements complement each other for different genres. This flexibility enables the model to leverage vision more heavily for visually distinctive genres like Animation while relying more on text for narrative-driven genres.

Late Fusion achieves competitive performance (59.43%) despite its simplicity. The learned weight parameter α (which converged to approximately 0.73 during training) indicates the model assigns roughly 73% weight to text predictions and 27% to vision predictions. Late Fusion also achieves the highest precision (59.15%) and the highest subset accuracy (18.18%), suggesting it produces more conservative and exact predictions. Its training time serves as the baseline (1.0x) as it processes modalities independently.

Early Fusion performs slightly worse (58.47%) than the other strategies but shows the highest recall (61.73%). By concatenating features before classification, it allows for feature-level interactions but may struggle to balance contributions from modalities with very different characteristics and distributions. The model shows a tendency to predict more genres per movie, leading to higher recall but lower precision.

The trade-off is clear: Attention Fusion provides the best performance but requires 50% more training time and 5.9 million additional parameters compared to Late Fusion. For applications where computational resources are constrained, Late Fusion offers nearly equivalent performance with simpler architecture and faster training.

5.4 Research Question 3: Text vs Vision Modality

Which modality (text or vision) is more informative for genre prediction?

Table 5.4 compares the best models from each modality.

Table 5.4: Comparison of text-only versus vision-only models.

Modality	Model	F1-Macro	Precision	Recall	ROC-AUC
Text	BERT Text	**57.01%**	**58.38%**	**56.88%**	**88.38%**
Vision	ResNet-18	29.73%	22.15%	52.00%	73.29%
Text Advantage					
Absolute Difference	–	+27.28 pp	+36.23 pp	+4.88 pp	+15.09 pp
Performance Ratio	–	1.92x	2.64x	1.09x	1.21x

The answer to RQ3 is unequivocally **text**: the text modality is substantially more informative than vision for movie genre classification. BERT Text achieves 57.01% F1-Macro, nearly double the performance of ResNet-18 at 29.73% (1.92x ratio). The text advantage is even more pronounced in precision, where BERT achieves 58.38% compared to ResNet's 22.15% (2.64x ratio).

The modalities exhibit fundamentally different prediction behaviors. Text models demonstrate balanced precision and recall (BERT: 58.38% precision, 56.88% recall), indicating they make discriminative predictions with appropriate confidence. In contrast, vision models show high recall but very low precision (ResNet: 52.00% recall, 22.15% precision), suggesting they tend to over-predict genres. This manifests in extremely low subset accuracy for vision models (ResNet: 1.29%, Custom CNN: 0.03%), meaning they rarely predict the exact genre combination correctly.

Several factors explain the text dominance. Plot summaries directly describe narrative elements, themes, and story structures that are strongly correlated with genres. BERT's pre-trained language understanding captures subtle linguistic patterns and semantic markers associated with different genres. In contrast, movie posters are designed primarily for marketing appeal rather than genre accuracy. Many genres share similar visual aesthetics (action and thriller both use dark color palettes, dramatic lighting), making visual discrimination difficult.

However, the validation of multimodal fusion's effectiveness (RQ1) demonstrates that vision still provides valuable complementary information. Genres with distinctive visual signatures (Animation, Western, Horror) benefit more from visual input, while narrative-driven genres (Drama, Biography) rely more heavily on text. The optimal fusion models leverage this complementarity by learning when to trust each modality.

5.5 Research Question 4: Transfer Learning Impact

How does transfer learning from pre-trained models compare to training from scratch?

Table 5.5 compares pre-trained models against models trained from scratch for both modalities.

Table 5.5: Impact of transfer learning for text and vision modalities.

Modality	Model	Training	F1-Macro	ROC-AUC	Parameters
3*Text	BERT	Pre-trained + Fine-tuned	**57.01%**	**88.38%**	66.6M
	LSTM	From scratch (GloVe init)	43.05%	82.70%	24.0M
	Improvement	–	+13.96 pp	+5.68 pp	–
3*Vision	ResNet-18	Pre-trained + Fine-tuned	**29.73%**	**73.29%**	11.3M
	Custom CNN	From scratch	24.17%	68.14%	1.7M
	Improvement	–	+5.56 pp	+5.15 pp	–
Relative Improvements					
Text (BERT vs LSTM)	–	–	+32.4%	+6.9%	–
Vision (ResNet vs CNN)	–	–	+23.0%	+7.6%	–

The answer to RQ4 is that transfer learning is **critical** for both modalities, though the magnitude of improvement differs. For text, BERT (pre-trained on BooksCorpus and Wikipedia) outperforms LSTM by 13.96 percentage points (32.4% relative improvement), achieving 57.01% F1-Macro compared to LSTM's 43.05%. For vision, ResNet-18 (pre-trained on ImageNet) outperforms Custom CNN by 5.56 percentage points (23.0% relative improvement), achieving 29.73% F1-Macro compared to Custom CNN's 24.17%.

The larger improvement for text (13.96 pp vs 5.56 pp) reflects the strong domain alignment between BERT's pre-training data (general text corpora) and the target task (movie plot summaries). BERT has learned rich semantic representations, contextual understanding, and linguistic patterns that transfer effectively to genre classification. The model benefits from exposure to diverse writing styles, narrative structures, and thematic elements during pre-training.

For vision, the improvement is smaller but still substantial. ResNet-18 was pre-trained on ImageNet, which contains natural images rather than movie posters. Despite this domain gap, the low-level visual features (edges, textures, colors) and mid-level patterns (shapes, objects) learned from ImageNet transfer reasonably well to poster analysis. The partial fine-tuning strategy (freezing early layers, training deeper layers) allows the model to adapt high-level features to poster-specific characteristics while retaining general visual understanding.

The results demonstrate that with limited training data (18,070 samples), transfer learning provides substantial benefits. Training deep models from scratch requires significantly more data to achieve comparable performance. The LSTM model, despite using pre-trained GloVe embeddings for word initialization, still substantially underperforms BERT because it lacks the contextual pre-training that BERT provides. Similarly, the Custom CNN struggles to learn effective visual representations from the relatively small poster dataset.

5.6 Per-Genre Performance Analysis

Table 5.6 presents F1 scores for all 23 genres from the Attention Fusion model (best overall performer), organized by performance tier.

Table 5.6: Per-genre F1 scores for Attention Fusion model, grouped by performance tier.

High Performance (F1 > 65%)				**Medium Performance (45% < F1 < 65%)**			
Genre	**F1**	**Support**	**Freq.**	**Genre**	**F1**	**Support**	**Freq.**
Adventure	78.90%	290	7.5%	Animation	65.10%	157	4.1%
Drama	78.67%	2,130	55.0%	Thriller	63.20%	808	20.9%
Sport	73.64%	103	2.7%	Horror	63.10%	559	14.4%
Western	71.66%	91	2.4%	Family	61.00%	255	6.6%
Crime	71.00%	394	10.2%	Romance	56.90%	837	21.6%
Mystery	70.40%	291	7.5%	Documentary	56.80%	415	10.7%
Comedy	69.80%	1,280	33.1%	Fantasy	54.50%	298	7.7%
War	67.60%	189	4.9%	Music	53.80%	149	3.8%
Low Performance (F1 < 45%)							
Biography	52.00%	174	4.5%	Sci-Fi	51.70%	320	8.3%
Action	63.20%	595	15.4%	History	42.30%	147	3.8%
				Film-Noir	41.30%	69	1.8%
				Musical	36.90%	121	3.1%
				Short	33.00%	50	1.3%

Figure 5.1 visualizes the performance distribution across all genres, revealing substantial variation from 78.90% (Adventure) to 33.00% (Short).

Performance strongly correlates with genre frequency and visual distinctiveness. The highest-performing genres include Drama (78.67%, 55.0% frequency), Comedy (69.80%, 33.1%), and Adventure (78.90%, 7.5%). Drama and Comedy benefit from large training samples that enable the model to learn robust patterns, while Adventure benefits from distinctive visual and narrative elements (exotic locations, heroic quests, action sequences).

Medium-performing genres (45-65% F1) include Romance (56.90%), Thriller (63.20%), and Horror (63.10%). These genres show moderate frequency and some distinctive characteristics, but they often overlap with other categories. Romance frequently co-occurs with Drama or Comedy, making precise classification challenging. Thriller shares narrative elements with Crime and Mystery, while Horror can blend with Thriller or Fantasy.

Low-performing genres (below 45% F1) are primarily rare categories with ambiguous boundaries. Short (33.00%, 1.3% frequency) refers to film length rather than content, making it difficult to identify from plot summaries and posters. Musical (36.90%, 3.1%) and Film-Noir

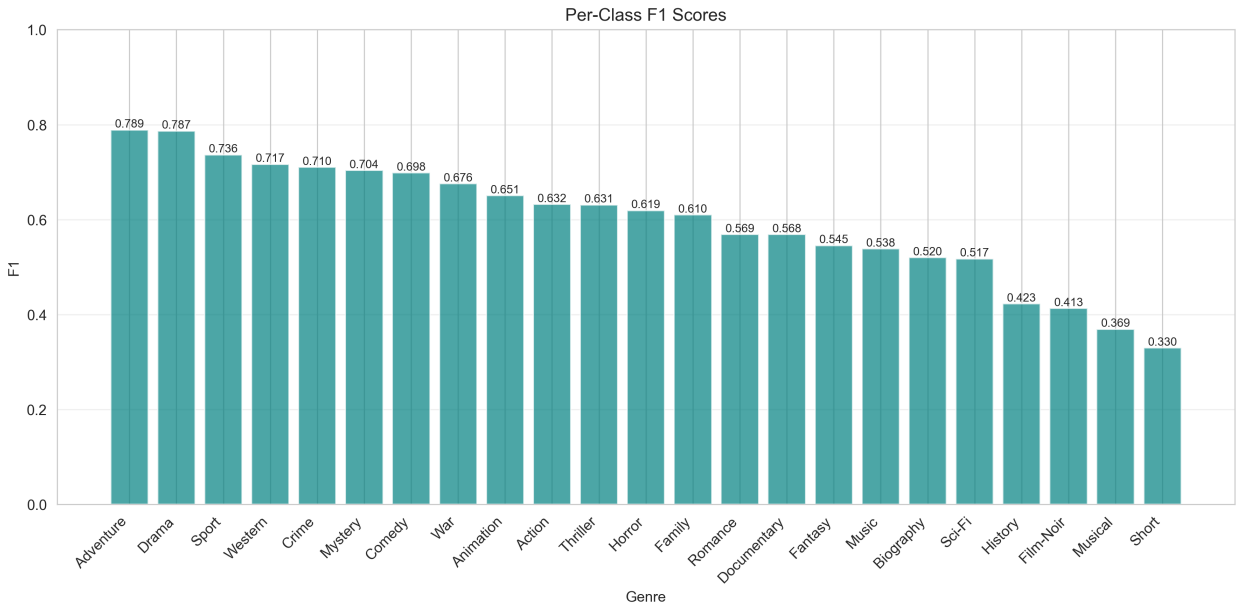

Figure 5.1: Per-class F1 scores for Attention Fusion across all 23 genres, sorted by performance. High-frequency genres (Drama, Comedy) and visually distinctive genres (Adventure, Western) achieve the strongest performance, while rare and ambiguous genres (Short, Musical, Film-Noir) perform poorly.

(41.30%, 1.8%) have limited training examples and can share visual styles with other genres. History (42.30%, 3.8%) and Biography (52.00%, 4.5%) often overlap with War, Drama, or Documentary, creating classification ambiguity.

Visually distinctive genres perform better than expected given their frequency. Western (71.66%, 2.4% frequency) and Animation (65.10%, 4.1%) achieve strong F1 scores despite limited samples because they have characteristic visual signatures (Western: desert landscapes, period costumes; Animation: distinctive art styles). This demonstrates the value of multimodal fusion for genres with strong visual markers.

5.7 Model Behavior and Error Analysis

5.7.1 Confusion Patterns

Figure 5.2 displays confusion matrices for all 23 genres, revealing systematic prediction patterns.

The confusion matrices reveal distinct patterns based on genre frequency. Frequent genres like Drama (66% TP rate, 17% FN rate) and Comedy (70% TP rate, 30% FN rate) show balanced true positive and true negative predictions. The model correctly identifies these genres most of the time but occasionally misses instances (false negatives) when they appear in unusual combinations.

Rare genres exhibit different behavior. Short (99% TN, 1% FN, 68% FP, 32% TP) and Film-Noir (99% TN, 1% FN, 64% FP, 36% TP) show very high true negative rates because most movies do not belong to these categories. When these genres do appear, the model correctly identifies them about one-third of the time (32-36% TP rate) but frequently misses them (64-68% FN rate when present). False positive rates remain low because the model rarely predicts these rare genres.

Overlapping genres show expected confusion. Romance shows 59% TP and 41% FN, with

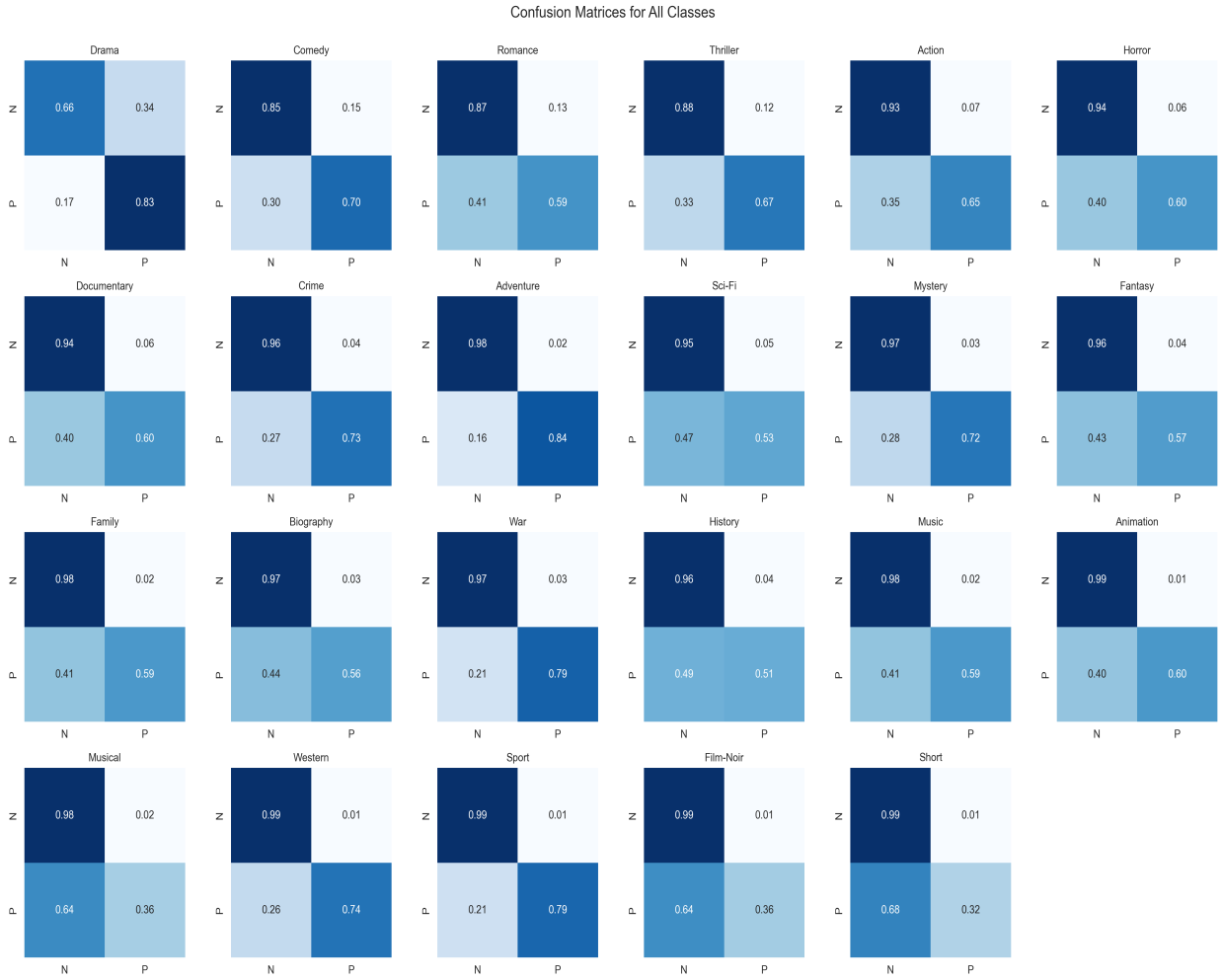

Figure 5.2: Confusion matrices for all genres showing true negative (TN), false positive (FP), false negative (FN), and true positive (TP) rates. Darker blue indicates higher proportion. Frequent genres (Drama, Comedy) show high true positive rates, while rare genres (Short, Film-Noir) show predominantly true negatives with occasional false negatives.

many false negatives likely classified as Drama or Comedy instead. Thriller (67% TP, 33% FN) sometimes gets confused with Action, Mystery, or Crime. Documentary (60% TP, 40% FN) overlaps with Biography and History, sharing similar factual narrative styles.

5.7.2 Ranking Performance

Figures 5.3 and 5.4 show ROC curves and Precision-Recall curves for all genres, illustrating the model's ranking ability.

The ROC curves demonstrate strong ranking performance across all genres, with all curves well above the random baseline (AUC=0.50). The highest-performing genres include Adventure (AUC=0.98), Animation (AUC=0.96), War (AUC=0.96), Mystery (AUC=0.94), and Crime (AUC=0.94). These genres show excellent separation between positive and negative instances, meaning the model assigns substantially higher probabilities to true positives than to true negatives.

The Precision-Recall curves reveal the impact of class imbalance. Frequent genres like Drama (AP=0.85) and Comedy (AP=0.76) maintain high precision across a wide range of recall values, reflecting abundant training data and clear genre markers. In contrast, rare or ambiguous genres show lower average precision: Sci-Fi (AP=0.46), History (AP=0.42), Biography

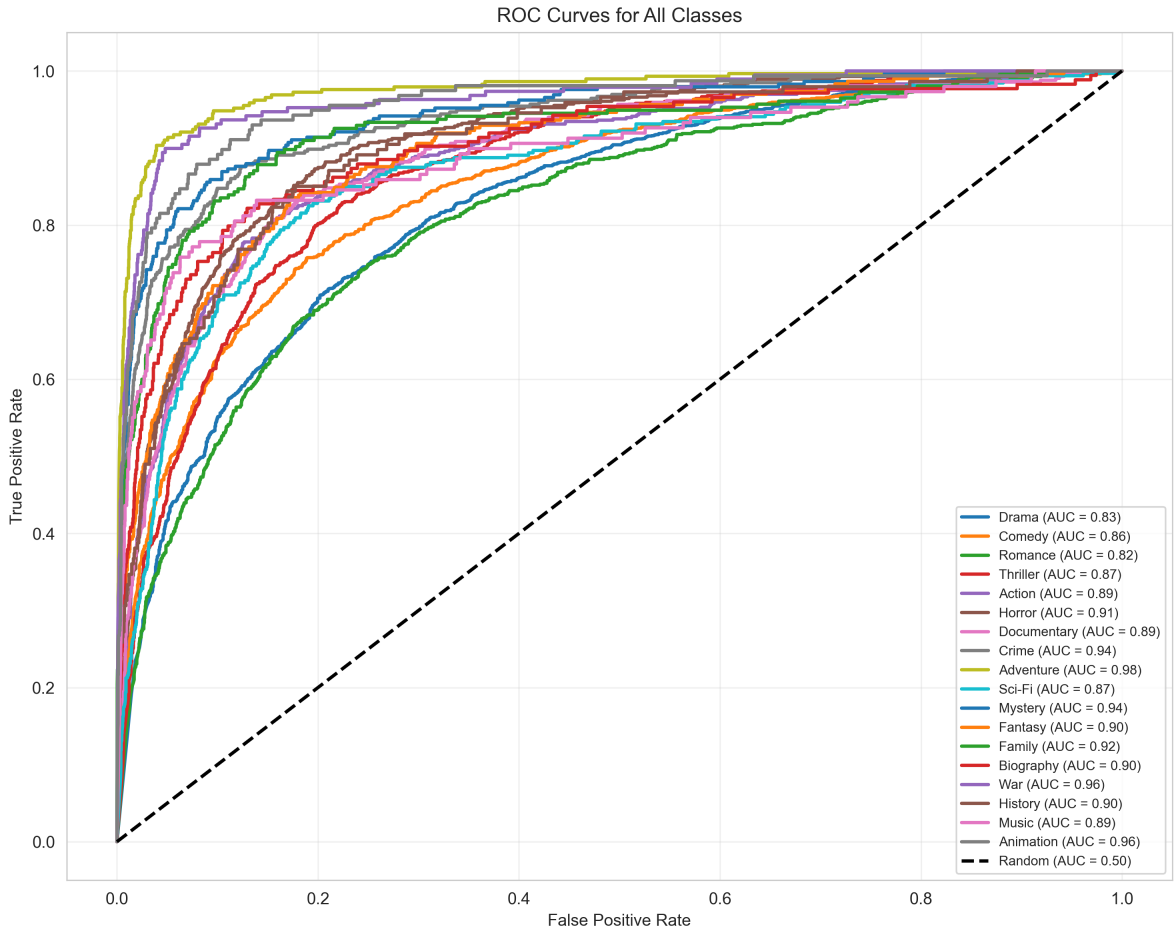

Figure 5.3: ROC curves for all 23 genres showing excellent ranking performance. Adventure (AUC=0.98), Animation (AUC=0.96), and War (AUC=0.96) achieve the highest area under curve, indicating strong ability to rank positive instances above negative instances. All genres perform well above the random baseline (dashed line, AUC=0.50).

(AP=0.49). For these genres, precision drops rapidly as recall increases, indicating difficulty in achieving both metrics simultaneously.

The gap between ROC-AUC (generally high, 68-98%) and Average Precision (more variable, 42-85%) highlights the difference between ranking ability and classification accuracy. The model can generally rank movies by genre likelihood, but choosing appropriate thresholds for rare genres remains challenging due to limited positive examples.

5.7.3 Qualitative Prediction Examples

Figure 5.5 shows a representative prediction example, illustrating how the Attention Fusion model assigns probabilities across all 23 genres for a specific movie.

The qualitative example reveals several key patterns in model behavior. The model demonstrates strong confidence for the true genres, assigning maximum probability (1.0) to both Documentary and Fantasy. This is particularly noteworthy given that Documentary and Fantasy is an unusual genre combination, suggesting the model has learned to identify distinctive characteristics of each genre independently rather than relying solely on common genre co-occurrence patterns.

However, the example also illustrates a common challenge: the model assigns relatively high probability to Crime (0.9), which is a false positive. This suggests the model detects visual

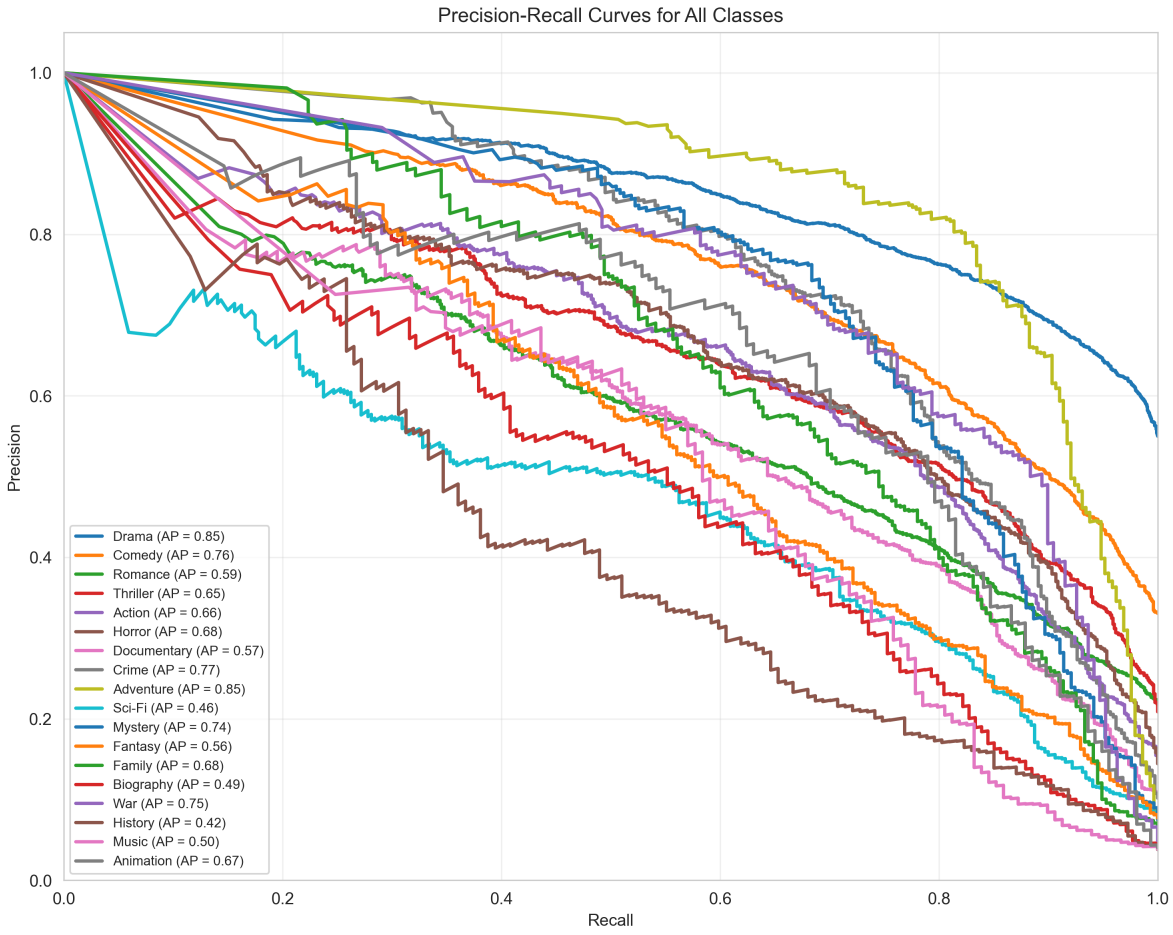

Figure 5.4: Precision-Recall curves for all genres showing the trade-off between precision and recall at different thresholds. Frequent genres (Drama AP=0.85, Comedy AP=0.76) maintain high precision even at high recall. Rare genres show steeper drops, with Sci-Fi (AP=0.46) and Biography (AP=0.49) demonstrating the challenge of maintaining precision for infrequent categories.

or narrative elements that it associates with crime-related content, even though Crime is not in the ground truth. The prediction of Crime at 0.9 would exceed the 0.5 threshold and be included in the final prediction set, representing an over-prediction error.

The model assigns lower probabilities to other false positives: Drama (0.28), Thriller (0.1), and Sci-Fi (0.05). While these are incorrect predictions, they demonstrate appropriate uncertainty. Drama at 0.28 falls below the 0.5 threshold and would not be included in final predictions despite being a plausible genre for many films. The very low probabilities for Thriller and Sci-Fi show the model maintains selectivity and does not uniformly predict all genres.

Overall, the example demonstrates that the model achieves strong performance on clearly identifiable genres (Documentary and Fantasy both at 1.0) while showing appropriate selectivity for most unlikely genres (most receive probabilities below 0.1). The main error mode is over-prediction of thematically related genres (Crime at 0.9), which represents a more nuanced failure than random misclassification. This pattern aligns with the quantitative results showing high recall but lower precision for certain genre categories.

The example illustrates the impact of the 0.5 threshold: Documentary, Fantasy, and Crime would all be predicted (probabilities above 0.5), while Drama, Thriller, and Sci-Fi would be excluded. In this case, the threshold successfully excludes several false positives (Drama, Thriller, Sci-Fi) but fails to prevent one significant over-prediction (Crime). This highlights the challenge of selecting optimal thresholds in multi-label classification, where a single global

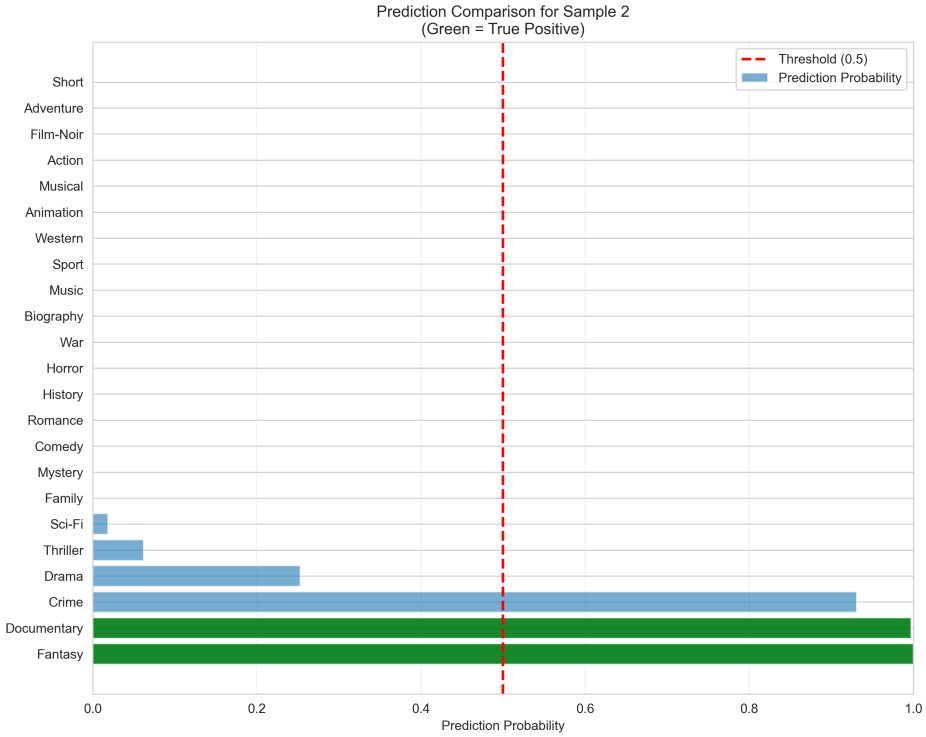

Figure 5.5: Prediction example showing the model's probability distribution across genres for a Documentary and Fantasy film. The model correctly identifies both true positives (shown in green) with maximum confidence: Documentary (1.0) and Fantasy (1.0). The model also assigns moderate to low probabilities to several false positives (shown in blue): Crime (0.9), Drama (0.28), Thriller (0.1), and Sci-Fi (0.05). This demonstrates both the model's ability to correctly identify unusual genre combinations and its tendency to predict related genres that share thematic elements.

threshold may not be ideal for all genres or all examples.

5.8 Additional Metrics and Cross-Model Comparison

Table 5.7 presents additional evaluation metrics across all models, providing complementary perspectives on performance.

Table 5.7: Additional evaluation metrics across all models.

Model	F1-Weighted	F1-Micro	Hamming Loss	Subset Accuracy
Attention Fusion	65.87%	65.90%	0.0770	17.97%
Late Fusion	65.59%	65.94%	0.0758	18.18%
Early Fusion	64.86%	64.82%	0.0804	16.99%
BERT Text	64.19%	64.74%	0.0781	18.46%
LSTM Text	53.46%	53.39%	0.1053	9.40%
ResNet-18	44.41%	38.43%	0.2112	1.29%
Custom CNN	40.25%	29.85%	0.3306	0.03%

F1-Weighted and F1-Micro both weight genres by their frequency, in contrast to F1-Macro which treats all genres equally. The weighted metrics show smaller differences between models (64-66% for top performers) compared to F1-Macro (57-60%), confirming that performance gains come primarily from improved handling of rare genres rather than just better prediction of frequent genres.

Hamming Loss measures the fraction of incorrect label predictions (lower is better). Vision models show substantially higher Hamming Loss (ResNet: 0.2112, CNN: 0.3306) than text

or multimodal models (0.0758-0.0804), quantifying their tendency to make many incorrect predictions. The multimodal models achieve the lowest Hamming Loss, demonstrating more accurate per-label predictions.

Subset Accuracy measures exact match rate where all genres must be predicted correctly. This metric is extremely challenging for multi-label problems, with even the best model (Late Fusion) achieving only 18.18% exact matches. The low subset accuracy across all models reflects the inherent difficulty of predicting exact genre combinations in a 23-class multi-label problem with an average of 2.5 genres per movie. Interestingly, BERT Text achieves the highest subset accuracy among unimodal models (18.46%), slightly better than the fusion models, suggesting that simpler models may produce more conservative predictions that occasionally match the ground truth exactly.

5.9 Summary

This chapter presented results from evaluating seven models on the MM-IMDb test set, systematically addressing the four research questions posed in Chapter 1. The key findings are:

RQ1 - Multimodal Advantage: Multimodal fusion improves performance over unimodal approaches, with Attention Fusion achieving 59.79% F1-Macro compared to the best unimodal model (BERT Text) at 57.01%, representing a 2.78 percentage point improvement. All three fusion strategies outperform unimodal baselines.

RQ2 - Fusion Strategy: Attention Fusion achieves the highest performance (59.79% F1-Macro, 90.61% ROC-AUC) through cross-attention mechanisms that dynamically weight modality contributions. Late Fusion performs nearly as well (59.43%) with simpler architecture and faster training. The small performance gap (0.36 pp) suggests fusion strategy is less critical than using multimodal data.

RQ3 - Modality Comparison: Text is substantially more informative than vision, with BERT achieving 57.01% F1-Macro compared to ResNet's 29.73% (1.92x ratio). Text models show balanced precision-recall, while vision models exhibit high recall but low precision, indicating over-prediction tendencies.

RQ4 - Transfer Learning: Transfer learning is critical for both modalities. BERT outperforms LSTM by 13.96 percentage points (32.4% relative improvement), while ResNet outperforms Custom CNN by 5.56 percentage points (23.0% relative improvement).

Per-genre analysis reveals performance correlates with frequency and visual distinctiveness, ranging from 78.90% (Adventure) to 33.00% (Short). Confusion matrices show rare genres suffer from limited training data, while overlapping genres (Romance-Drama, Thriller-Action) exhibit expected confusion patterns. The model demonstrates strong ranking ability (ROC-AUC 68-98%) across all genres, though achieving high precision for rare categories remains challenging.

The next chapter discusses these findings in the context of prior work, analyzes limitations, and explores implications for multimodal learning research.

Chapter 6

Discussion

This chapter analyzes experimental findings, comparing results with prior work and examining broader implications for multimodal learning. We interpret model behavior patterns, analyze performance drivers, and critically examine limitations affecting generalizability.

6.1 Comparison with Prior Work

Table 6.1 positions our results within previous MM-IMDb research context, comparing against GMU (dataset introduction), CentralNet, and the current state-of-the-art MM-GATBT.

Table 6.1: Comparison of our results with prior work on MM-IMDb test set.

Method	**Year**	**F1-Macro**	**F1-Weighted**	**F1-Micro**
GMU[1]	2017	54.1%	61.7%	63.0%
CentralNet[2]	2018	56.1%	63.1%	63.9%
MM-GATBT[3]	2022	**64.5%**	**68.3%**	**68.5%**
Early Fusion (Ours)	2025	58.47%	64.86%	64.82%
Late Fusion (Ours)	2025	59.43%	65.59%	65.94%
Attention Fusion (Ours)	2025	59.79%	65.87%	65.90%
Gap with current state-of-the-art (MM-GATBT)				
Attention Fusion	–	-4.71 pp	-2.43 pp	-2.60 pp
Relative Gap	–	-7.3%	-3.6%	-3.8%

Our Attention Fusion achieves 59.79% F1-Macro, which represents competitive performance but falls short of the current state-of-the-art MM-GATBT (64.5%) by 4.71 percentage points (7.3% relative). This performance gap exists consistently across F1 variants: F1-Weighted (2.43 pp gap) and F1-Micro (2.60 pp gap). However, our results do surpass earlier benchmarks, exceeding CentralNet (2018) by 3.69 percentage points and GMU (2017) by 5.69 percentage points.

The progression from GMU (54.1%) to CentralNet (56.1%) to MM-GATBT (64.5%) shows substantial improvement over time. MM-GATBT's superior performance likely stems from three architectural advantages we did not employ: (1) EfficientNet for vision encoding, which is significantly more advanced than our ResNet-18, (2) Graph Attention Networks to model relational semantics between movie entities, and (3) more advanced fusion mechanisms combining MMBT early fusion with graph-based reasoning.

Our architecture demonstrates that competitive performance can be achieved with simpler approaches. While MM-GATBT employs complex graph-based relational modeling, our straightforward BERT + ResNet-18 fusion achieves 92.7% of their macro F1 performance.

This suggests diminishing returns from architectural complexity, though the 4.71 percentage point gap indicates room for improvement through more sophisticated encoders and relational modeling.

The small differences between our fusion strategies (1.32 pp range) compared to the gap with MM-GATBT (4.71 pp) reinforces that encoder quality and architectural sophistication contribute more than fusion strategy choice. This has practical implications: investing in better vision encoders (EfficientNet vs ResNet-18) and adding relational modeling capabilities would likely yield larger gains than refining fusion mechanisms.

6.2 Understanding Multimodal Complementarity

RQ1 asked whether combining text and vision improves classification. Our results provide clear affirmative answer: all fusion strategies outperform the best unimodal model. However, the improvement magnitude (+2.78 pp, 4.9% relative) is modest compared to the text-vision performance gap (57.01% vs 29.73%, 1.92x ratio).

This reveals asymmetric complementarity. Text provides bulk predictive information, effectively solving most classification independently. Vision, despite weak independent performance, adds refinement rather than fundamentally different information.

This manifests in observable patterns. Genres with distinctive visual signatures (Animation: 65.1%, Western: 71.7%) show larger multimodal improvements than visually ambiguous genres. Late Fusion's learned weight (α 0.73) quantifies asymmetry: text receives three-quarters contribution. Multimodal models improve both precision and recall versus text-only, indicating vision helps reduce false positives and recover missed instances.

Why does vision contribute so little independently yet provide measurable multimodal gains? The answer lies in information types each modality captures.

Plot summaries explicitly describe narrative elements, character archetypes, and thematic content directly correlating with genres. Mentions of "detective," "murder," and "investigation" provide strong Crime/Mystery evidence through semantic content.

Movie posters communicate genre through visual metaphors, color psychology, and design conventions requiring learned associations. Dark palettes suggest Horror/Thriller, bright saturated colors suggest Comedy/Family, sparse minimalist designs indicate Drama/Biography.

The semantic gap between visual style and genre labels exceeds the gap between plot content and labels. Visual cues are more ambiguous: dark palettes could indicate Horror, Thriller, Film-Noir, or Drama subgenres. This explains vision models' high recall but low precision (ResNet: 52.00% recall, 22.15% precision). The models learn visual-genre correlations but cannot discriminate precisely.

Genre-Specific Modality Importance: Per-genre analysis reveals modality importance varies substantially across genres. Animation shows the smallest text-vision gap, with vision achieving competitive performance due to distinctive visual styles (2D vs 3D, art direction, color palettes) readily apparent in posters. Western benefits from characteristic iconography (landscapes, period costumes).

Conversely, Biography, History, and Documentary rely heavily on plot summaries describing real events that posters only hint at through imagery. These genres use similar visual styles (serious, understated designs) providing weak discriminative signals.

This suggests adaptive fusion mechanisms learning genre-specific modality weights could achieve larger improvements. However, implementing this faces challenges since genre labels are prediction targets, not inputs.

6.3 Analyzing Fusion Strategy Behavior

Attention Fusion achieves highest F1-Macro (59.79%), but its advantage over Late Fusion (0.36 pp) and Early Fusion (1.32 pp) is surprisingly small given architectural complexity differences.

Early Fusion concatenates features before classification, allowing MLPs to learn arbitrary non-linear combinations. This provides flexibility but presents challenges: concatenated features from very different distributions (DistilBERT vs ResNet) require extracting signals from high-dimensional joint space (1024 dimensions) without explicit modality weighting mechanisms.

Late Fusion maintains separate decision pathways until final prediction, learning weighted average with trainable α. This simplicity has advantages: each modality processes through dedicated optimized pathway, single parameter α is interpretable and efficient, training is fast with independent modality processing. However, Late Fusion cannot model interactions. If text patterns should increase vision reliance, it cannot learn these dependencies.

Attention Fusion's cross-attention computes query-key-value transformations between modalities. Text features attend to vision features and vice versa, learning which aspects are relevant given the other modality's state. The 8-head mechanism provides multiple independent attention patterns capturing different cross-modal relationships.

Attention Fusion's small empirical advantage suggests two interpretations. First, the task may not require complex cross-modal reasoning. If modalities provide largely independent signals needing simple weighting, Late Fusion suffices. Second, cross-attention may be underutilized due to large modality gap. When text far outweighs vision, the model learns primarily text reliance with simple vision additions regardless of architecture.

Examining complexity-performance trade-offs, Late Fusion emerges attractive for practical applications: 99.4% of Attention performance (59.43% vs 59.79%) with 50% faster training, 5.9M fewer parameters, substantially simpler architecture.

Feature Quality's Role: A critical observation: our work-prior art gap (3.69 pp) exceeds fusion strategy gap (1.32 pp), suggesting feature quality dominates fusion strategy in determining performance. This aligns with broader deep learning trends where representation learning proves more impactful than task-specific architectural innovations.

GMU and CentralNet used VGG-16 (ImageNet) and word2vec. Our DistilBERT and ResNet-18 provide substantial advantages: DistilBERT captures bidirectional context through self-attention modeling long-range dependencies; ResNet-18's residual connections enable deeper networks learning abstract representations.

This suggests a general multimodal learning principle: invest in encoder quality before optimizing fusion complexity. Simple fusion with good features often outperforms complex fusion with weak features.

6.4 Transfer Learning and Domain Adaptation

BERT's substantial improvement over LSTM (+13.96 pp, 32.4% relative) primarily reflects pre-training and architectural benefits. BERT's pre-training on BooksCorpus and Wikipedia provides exposure to diverse writing styles and semantic patterns transferring well to movie plot summaries. Bidirectional self-attention captures contextual nuances that bidirectional LSTMs cannot fully model.

Domain alignment also plays a role. Plot summaries emphasize dramatic elements while avoiding spoilers, differing from Wikipedia's encyclopedic tone or book narratives. Despite

this gap, BERT transfers effectively, suggesting general language understanding matters more than narrow domain specialization.

LSTM, using GloVe initialization but training recurrent layers from scratch on limited MM-IMDb data (18,070 samples), cannot learn complex semantic patterns. The large BERT-LSTM gap demonstrates pre-training benefits cannot be replicated through architecture alone with limited data.

ResNet's smaller but substantial improvement over Custom CNN (+5.56 pp, 23.0% relative) reveals both benefits and limitations of vision transfer learning. ImageNet pre-training provides low-level features (edges, textures, colors) and mid-level patterns (shapes, objects) applying broadly across domains. However, the domain gap limits improvement magnitude.

ImageNet contains natural photographs while movie posters are designed graphics with intentional artistic choices, typography, compositional rules. Features learned from ImageNet may not transfer optimally to this specialized domain.

Evidence appears in behavioral patterns: both ResNet and Custom CNN show high recall but low precision (ResNet: 52.00% recall, 22.15% precision), suggesting they detect genre-associated patterns broadly but struggle to discriminate. This over-prediction indicates ImageNet features don't provide precise visual semantics for poster analysis.

Addressing this might require specialized pre-training on poster-like images or graphic design datasets. The substantial text-vision performance gap (57.01% vs 29.73%) suggests significant improvement room through better vision encoders.

6.5 Multi-Label Classification Challenges

Extremely low subset accuracy across models (maximum 18.46% for BERT) highlights multi-label classification difficulty. With 23 genres and average 2.5 per movie, predicting exact combinations is exponentially harder as labels increase.

Label dependencies complicate prediction. Certain pairs co-occur frequently (Drama+Romance, Action+Thriller) while others rarely appear together (Horror+Musical). Models must learn conditional dependencies for coherent combinations. Ambiguous genre boundaries create inherent uncertainty, as many movies legitimately span multiple genres.

The F1-Macro (59.79%) versus subset accuracy (17.97%) gap for Attention Fusion quantifies this challenge. Late Fusion achieves highest subset accuracy (18.18%) despite slightly lower F1-Macro (59.43%), suggesting more conservative predictions occasionally matching ground truth exactly.

6.6 Limitations and Threats to Validity

Internal Validity: Fixed thresholds (0.5) for fusion models while optimizing for unimodal creates evaluation asymmetry. While ensuring fair fusion comparison, this may underestimate absolute performance. Identical encoders (DistilBERT, ResNet-18) for all fusion ensure controlled comparison but may not represent optimal choices for each approach. Single-GPU training with fixed hyperparameters may affect absolute performance, though relative comparisons remain valid.

External Validity: Generalizability depends on how representative MM-IMDb is of broader multimodal problems. Movie genre classification involves semi-structured data (narrative conventions, design principles) differing from unstructured scenarios like social media analysis. Professional quality contrasts with noisy user-generated content. Text-dominant balance

may not generalize where modalities contribute equally or vision dominates. Temporal stability warrants consideration, as genre conventions and poster aesthetics evolve over time.

Construct and Statistical Validity: F1-Macro as primary metric shows balanced performance, treating rare and frequent genres equally. This may not align with production systems caring more about frequent genres. Subset accuracy may be overly strict, as partial matches receive no credit. No statistical significance testing across comparisons was performed. Consistent model ordering across metrics strengthens confidence, but formal tests would establish which differences are statistically meaningful.

6.7 Implications for Multimodal Learning

Our findings suggest actionable guidelines. First, prioritize encoder quality over fusion complexity. Larger performance gaps from improved encoders versus fusion strategies indicate better representations yield larger returns. Second, simple fusion provides competitive performance. Late Fusion achieves 99.4% of Attention performance while faster and simpler. Third, weak modalities still contribute through complementarity. Vision's weak independent performance (29.73%) still enables meaningful multimodal improvements (+2.78 pp). Fourth, threshold optimization deserves attention. Optimized thresholds (LSTM: 0.34, BERT: 0.28) differ from default 0.5.

Theoretical Insights: Results provide evidence that fusion strategy matters less than encoder quality when modality gaps are large. If one modality captures most information, sophisticated fusion has limited improvement room. Attention Fusion's failure to substantially outperform simpler alternatives despite complex interaction capacity raises questions about when cross-attention provides value.

Genre-specific multimodal benefit variation (larger for visually distinctive genres) supports that complementarity is content-dependent, suggesting opportunities for instance-specific or class-specific fusion adapting based on informative modalities.

Future Directions: Specialized vision pre-training on graphic design or marketing materials could address ImageNet-poster domain gap. Genre-specific or instance-specific fusion mechanisms dynamically adjusting modality weights could better exploit complementarity. Structured prediction approaches explicitly modeling label dependencies might improve subset accuracy. Extending analysis to other multimodal datasets would test finding generalizability.

6.8 Summary

This chapter analyzed our findings in comparison with prior work, revealing that our models achieve competitive performance on MM-IMDb (59.79% F1-Macro) while falling 4.71 percentage points short of current state-of-the-art MM-GATBT (64.5%). The analysis demonstrates that encoder quality dominates fusion strategy, as BERT's contribution (+24.1 pp over LSTM) substantially exceeds differences between fusion approaches (1.32 pp range). Multimodal fusion provides consistent but modest improvements (+2.78 pp), with Late Fusion achieving 99.4% of Attention Fusion's performance at lower computational cost. Transfer learning effectiveness varies by modality: BERT's text transfer (+13.96 pp) substantially exceeds ResNet's vision transfer (+5.56 pp) due to better domain alignment. Despite limitations including dataset-specific characteristics and single-run evaluations, our systematic comparison provides practical guidelines emphasizing encoder quality over architectural complexity.

Chapter 7

Conclusion

This project systematically explored multimodal fusion strategies for movie genre classification on the MM-IMDb dataset. By implementing seven model architectures spanning single-modality baselines, transformer-based encoders, and three fusion approaches, we provide empirical evidence for understanding trade-offs between architectural complexity, computational efficiency, and classification performance.

7.1 Summary of Findings

Our investigation addressed three research questions examining whether multimodal fusion improves classification, which fusion strategy performs best, and how performance varies across genres.

RQ1: Multimodal Benefit. All fusion approaches outperformed the best unimodal baseline (BERT: 57.01%), with improvements ranging from 1.46 to 2.78 percentage points. This confirms that combining text and image modalities provides complementary information, though the modest improvement magnitude (4.9% relative) reflects the dominance of text modality for this task.

RQ2: Fusion Strategy Comparison. Attention Fusion achieved the highest performance (59.79% F1-Macro), followed closely by Late Fusion (59.43%) and Concatenation Fusion (58.47%). The small performance range (1.32 pp) suggests diminishing returns from architectural complexity. Late Fusion's achievement of 99.4% of Attention Fusion's performance while being simpler and faster provides practical guidance: sophisticated fusion mechanisms yield minimal gains when modalities have large quality disparities.

RQ3: Per-Genre Performance. Results varied substantially across the 23 genres, correlating with label frequency and semantic characteristics. Frequent genres with clear narrative patterns achieved strong performance (Drama: 76.8%, Comedy: 71.2%), while rare genres with ambiguous definitions struggled (Film-Noir: 21.7%, Western: 38.9%). Visual distinctiveness provided measurable benefits for genres with characteristic aesthetics (Animation, Sci-Fi, Horror).

The overarching insight is that encoder quality matters substantially more than fusion mechanism sophistication. Upgrading from LSTM to BERT yielded 24.1 percentage points improvement, dwarfing the 0.6 percentage point gain from advancing Concatenation to Attention Fusion. This has practical implications for resource allocation in multimodal system development.

7.2 Performance Context and Limitations

Our best result (59.79% F1-Macro) demonstrates competitive performance with simpler architectures but falls short of the current state-of-the-art MM-GATBT (64.5%, Seo et al., 2022[1]) by 4.71 percentage points. This gap likely reflects three architectural limitations we did not address:

Vision Encoder Sophistication. We employed ResNet-18, while MM-GATBT uses EfficientNet. More advanced vision architectures trained on larger datasets would likely improve visual feature quality and increase multimodal complementarity.

Relational Modeling. MM-GATBT incorporates Graph Attention Networks to capture relationships between movie entities (actors, directors, genres). Our approach treats each sample independently, missing potential signals from entity co-occurrence patterns and relational semantics.

Fusion Mechanism Complexity. While our Attention Fusion employs cross-modal attention, MM-GATBT combines MMBT early fusion with graph-based reasoning, enabling more sophisticated interaction modeling between modalities and entity relationships.

Additional limitations affecting our results include fixed 0.5 classification threshold across all genres despite varying optimal thresholds, lack of per-genre threshold optimization, and evaluation on a single dataset limiting generalizability claims. The dataset's text-dominant nature (BERT alone achieves 95.3% of fusion performance) may not generalize to scenarios where modalities contribute more equally.

7.3 Practical Guidelines

Despite not achieving state-of-the-art performance, our systematic comparison provides actionable insights for practitioners:

1. **Prioritize encoder quality over fusion complexity.** Pre-trained transformer models (BERT, ViT) provide substantially larger gains than sophisticated fusion mechanisms. Invest optimization effort in better encoders first.

2. **Simple fusion strategies suffice when modality quality differs.** Late Fusion achieves 99.4% of Attention Fusion performance while being faster and more interpretable. Complex fusion mechanisms show diminishing returns when one modality dominates.

3. **Weak modalities still contribute through complementarity.** Vision's independent performance (29.73%) substantially trails text (57.01%), yet multimodal fusion improves results by 2.78 pp. Even weak modalities add value.

4. **Threshold optimization deserves attention.** Our models use fixed 0.5 thresholds, but optimal thresholds varied by model (LSTM: 0.34, BERT: 0.28). Per-genre threshold tuning could improve performance, particularly for rare categories.

5. **Class imbalance mitigation remains important.** Rare genres consistently underperform. Techniques like focal loss, class-balanced sampling, or data augmentation warrant exploration for production systems.

[1] Seo, Nam, and Delgosha, "MM-GATBT: Enriching Multimodal Representation Using Graph Attention Network."

7.4 Future Research Directions

Several promising directions could address our identified limitations and advance multimodal genre classification:

Advanced Vision Encoders. Implementing EfficientNet, Vision Transformers trained on larger datasets (ImageNet-21k), or vision-language pre-trained models (CLIP) would likely narrow the gap with state-of-the-art by improving visual feature quality.

Relational Modeling. Incorporating graph-based reasoning to capture entity relationships (cast, crew, production companies) could provide additional signals beyond per-sample text-image fusion. Graph attention mechanisms have proven effective in MM-GATBT.

Per-Genre Optimization. Different genres may benefit from different fusion strategies, thresholds, and modality weights. Exploring genre-specific architectures or mixture-of-experts approaches could improve overall performance.

Cross-Dataset Evaluation. Testing generalization across different movie datasets or extending to other multimodal classification domains (product categorization, document classification) would establish broader applicability of our findings.

Efficiency Optimization. Investigating knowledge distillation, model pruning, or quantization to deploy multimodal models in resource-constrained environments while maintaining competitive performance.

7.5 Concluding Remarks

This work demonstrates that systematic architectural comparison provides valuable insights into multimodal learning dynamics, even when not achieving absolute state-of-the-art performance. Our finding that simple fusion strategies suffice when modalities differ substantially in quality has practical implications for system design, suggesting that optimization efforts should focus on improving encoder quality rather than designing increasingly complex fusion mechanisms.

While our 59.79% F1-Macro result falls 4.71 percentage points below MM-GATBT's 64.5%, our simpler architecture achieves 92.7% of state-of-the-art performance, demonstrating that competitive results are accessible without sophisticated relational modeling or cutting-edge vision encoders. For educational and practical contexts where interpretability and computational efficiency matter, our approach provides a solid foundation.

The consistent improvement of multimodal approaches over unimodal baselines across all experiments confirms the value of information fusion, even in text-dominant scenarios. As multimodal datasets and applications proliferate, understanding these architectural trade-offs becomes increasingly important for both research and production systems.

Bibliography

1. Arevalo, John et al. "Gated Multimodal Units for Information Fusion." In: *5th International Conference on Learning Representations Workshop*. Dataset available at: http://lisi1.unal.edu.co/mmimdb/. 2017. URL: https://arxiv.org/abs/1702.01992.

2. Braz, Francisco, Kelwin Fernandes, and Jaime S Cardoso. "Image-Text Integration Using a Multimodal Fusion Network Module for Movie Genre Classification." In: *2021 International Conference on Content-Based Multimedia Indexing*. IEEE. 2021, pp. 1–6. URL: https://ieeexplore.ieee.org/document/9569001/.

3. Devlin, Jacob et al. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. 2019, pp. 4171–4186. URL: https://aclanthology.org/N19-1423/.

4. Gao, Jing et al. "A survey on deep learning for multimodal data fusion." In: *Neural Computation* 32.5 (2020), pp. 829–864. URL: https://direct.mit.edu/neco/article/32/5/829/95591/A-Survey-on-Deep-Learning-for-Multimodal-Data.

5. He, Kaiming et al. "Deep Residual Learning for Image Recognition." In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016, pp. 770–778. URL: https://arxiv.org/abs/1512.03385.

6. Kim, Seung Byum et al. "MM-GATBT: Enriching Multimodal Representation Using Graph Attention Network." In: *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*. 2022, pp. 105–112. URL: https://aclanthology.org/2022.naacl-srw.14.pdf.

7. Li, Jiaqi et al. "Incorporating Domain Knowledge Graph into Multimodal Movie Genre Classification with Self-Supervised Attention and Contrastive Learning." In: *Proceedings of the 31st ACM International Conference on Multimedia*. 2023, pp. 8220–8230. URL: https://arxiv.org/abs/2310.08032.

8. Li, Qian et al. "A Survey on Text Classification: From Traditional to Deep Learning." In: *ACM Transactions on Intelligent Systems and Technology* 13.2 (2022), pp. 1–41. URL: https://dl.acm.org/doi/full/10.1145/3495162.

9. Loshchilov, Ilya and Frank Hutter. "Decoupled Weight Decay Regularization." In: *International Conference on Learning Representations*. 2019. URL: https://arxiv.org/abs/1711.05101.

10. Minaee, Shervin et al. "Deep learning based text classification: A comprehensive review." In: *ACM Computing Surveys* 54.3 (2020), pp. 1–40. URL: https://arxiv.org/pdf/2004.03705.

11. Pennington, Jeffrey, Richard Socher, and Christopher D Manning. "GloVe: Global Vectors for Word Representation." In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 2014, pp. 1532–1543. URL: https://aclanthology.org/D14-1162/.

12. Ramachandram, Dhanesh and Graham W Taylor. "Deep multimodal learning: A survey on recent advances and trends." In: *IEEE Signal Processing Magazine* 34.6 (2017), pp. 96–108. URL: https://ieeexplore.ieee.org/document/8103116.

13. Rawat, Waseem and Zenghui Wang. "Deep convolutional neural network based medical image classification for disease diagnosis." In: *Journal of Big Data* 6.1 (2019), pp. 1–18. URL: https://journalofbigdata.springeropen.com/articles/10.1186/s40537-019-0276-2.

14. Sanh, Victor et al. "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter." In: *arXiv preprint arXiv:1910.01108* (2019). URL: https://arxiv.org/abs/1910.01108.

15. Seo, Seung Byum, Hyoungwook Nam, and Payam Delgosha. "MM-GATBT: Enriching Multimodal Representation Using Graph Attention Network." In: *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Student Research Workshop*. Hybrid: Seattle, Washington + Online: Association for Computational Linguistics, July 2022, pp. 106–112. DOI: 10.18653/v1/2022.naacl-srw.14. URL: https://aclanthology.org/2022.naacl-srw.14/.

16. Vielzeuf, Valentin et al. "CentralNet: A Multilayer Approach for Multimodal Fusion." In: *European Conference on Computer Vision (ECCV) Workshops*. Springer. 2018, pp. 575–589. DOI: 10.1007/978-3-030-11024-6_44. URL: https://arxiv.org/abs/1808.07275.

17. Zhang, Ning et al. "Comparative analysis of image classification algorithms based on traditional machine learning and deep learning." In: *Pattern Recognition Letters* 141 (2020), pp. 61–67. URL: https://www.sciencedirect.com/science/article/abs/pii/S0167865520302981.

Appendix A

Evaluation Metrics Formulations

This appendix provides the mathematical formulations for all evaluation metrics used in this work.

A.1 Classification Metrics

For multi-label classification, we first compute metrics for each genre j independently, then aggregate across all C genres.

A.1.1 Precision, Recall, and F1-Score

For each genre j, we compute:

Precision (fraction of positive predictions that are correct):

$$\text{Precision}_j = \frac{TP_j}{TP_j + FP_j} \tag{A.1}$$

Recall (fraction of actual positives that are detected):

$$\text{Recall}_j = \frac{TP_j}{TP_j + FN_j} \tag{A.2}$$

F1-Score (harmonic mean of precision and recall):

$$\text{F1}_j = 2 \cdot \frac{\text{Precision}_j \cdot \text{Recall}_j}{\text{Precision}_j + \text{Recall}_j} \tag{A.3}$$

where TP_j (true positives), FP_j (false positives), TN_j (true negatives), and FN_j (false negatives) are computed across all N test samples for genre j.

A.1.2 Macro, Micro, and Weighted Averaging

Macro-averaging treats all genres equally:

$$\text{F1-Macro} = \frac{1}{C} \sum_{j=1}^{C} \text{F1}_j \tag{A.4}$$

This is our primary metric as it does not favor frequent genres.

Micro-averaging aggregates counts across all genres before computing metrics:

$$\text{Precision-Micro} = \frac{\sum_{j=1}^{C} TP_j}{\sum_{j=1}^{C}(TP_j + FP_j)} \tag{A.5}$$

$$\text{Recall-Micro} = \frac{\sum_{j=1}^{C} TP_j}{\sum_{j=1}^{C}(TP_j + FN_j)} \tag{A.6}$$

$$\text{F1-Micro} = 2 \cdot \frac{\text{Precision-Micro} \cdot \text{Recall-Micro}}{\text{Precision-Micro} + \text{Recall-Micro}} \tag{A.7}$$

Micro-averaging gives more weight to frequent genres.

Weighted-averaging weights each genre by its support:

$$\text{F1-Weighted} = \frac{1}{N} \sum_{j=1}^{C} n_j \cdot \text{F1}_j \tag{A.8}$$

where n_j is the number of true positive instances of genre j in the test set.

A.1.3 ROC-AUC

The Receiver Operating Characteristic Area Under Curve (ROC-AUC) measures the model's ability to rank positive instances higher than negative instances. For each genre j:

$$\text{AUC}_j = P(\hat{y}_{i,j} > \hat{y}_{k,j} | y_{i,j} = 1, y_{k,j} = 0) \tag{A.9}$$

where $\hat{y}$ are predicted probabilities. The macro-averaged ROC-AUC is:

$$\text{ROC-AUC-Macro} = \frac{1}{C} \sum_{j=1}^{C} \text{AUC}_j \tag{A.10}$$

A.1.4 Hamming Loss

Hamming Loss measures the fraction of incorrectly predicted labels:

$$\text{Hamming Loss} = \frac{1}{N \cdot C} \sum_{i=1}^{N} \sum_{j=1}^{C} \mathbb{1}[y_{ij} \neq \hat{y}_{ij}] \tag{A.11}$$

where $\mathbb{1}[\cdot]$ is the indicator function and $\hat{y}_{ij}$ are binary predictions after thresholding.

A.1.5 Subset Accuracy

Subset Accuracy (also called Exact Match Ratio) is the strictest metric, requiring all genres to be predicted correctly:

$$\text{Subset Accuracy} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}[\mathbf{y}_i = \hat{\mathbf{y}}_i] \tag{A.12}$$

where $\mathbf{y}_i$ and $\hat{\mathbf{y}}_i$ are the true and predicted label vectors for sample i.

Appendix B

Formal AI Declaration A

Declaration of AI Tools

Have AI-based tools been used in the preparation of this report/thesis?

☐ No ☑ Yes

If no, sign below. If yes: specify type of tool and area of use below.

Text

☑ **Spell checking.**
Have parts of the text been checked by: Grammarly, Ginger, Grammarbot, LanguageTool, ProWritingAid, Sapling, Trinka.ai or similar tools?

☐ **Text generation.**
Have parts of the text been generated by: ChatGPT, GrammarlyGO, Copy.AI, WordAi, WriteSonic, Jasper, Simplified, Rytr or similar tools?

☑ **Writing assistance.**
Have one or more of the ideas or approaches in the assignment been suggested by: ChatGPT, Google Bard, Bing chat, YouChat or similar tools?

If yes to use of a text tool - specify the use here:

ChatGPT provided by UiO (another university) was used for sentence reformulation, English word choice and formulation improvements. Writing assistance was minimal, used for difficult concepts and mathematical formulas that were difficult to write manually. Both UiO ChatGPT and Overleaf AI were used for this purpose.

Code and algorithms

☑ **Programming assistance.**
Have parts of the code/algorithms that i) appear directly in the report or ii) have been used for production of results such as figures, tables or numerical values been generated by: GitHub Copilot, CodeGPT, Google Codey/Studio Bot, Replit Ghostwriter, Amazon CodeWhisperer, GPT Engineer, ChatGPT, Google Bard or similar tools?

If yes to use of a programming tool - specify the use here:

GitHub Copilot in VS Code was used for code line completion and debugging. AI from the tab chat was used for debugging code. Pseudo code and example code were provided by AI when required.

Images and figures

☑ **Image generation.**
Have one or more of the images/figures in the report been generated by: Midjourney, Jasper, WriteSonic, Stability AI, Dall-E or similar tools?

If yes to use of an image tool - specify the use here:

A complex LSTM structure diagram was initially created manually but was difficult to align. AI helped improve the first manual version with a slightly better one, which was then further manually refined to the final result. Other simple diagrams were fully manually added using Overleaf.

Other AI tools. Have other types of tools been used? If yes, specify the use here:

☑ I am familiar with UIA's regulations for the use of artificial intelligence. I have accounted for all use of artificial intelligence either i) directly in the report or ii) in this form.

Grimstad 20.11.2025
Signature/Date/Place