Journal of Advances in Engineering Sciences and Technology, 2025, 1(1), 17-34
https://doi.org/10.71204/rxht1w29

A Multimodal Deep Learning Framework for Video-Based Sentiment Analysis

Jinge Bai [1], Ainuddin Wahid Bin Abdul Wahab [1, *]

[1] Faculty of Computer Science and Information Technology, Universiti Malaya, Kuala Lumpur 50603, Malaysia

*** Correspondence:**

Ainuddin Wahid Bin Abdul Wahab

ainuddin@um.edu.my

Received: 30 April 2025/ Accepted: 18 October 2025/ Published online: 27 October 2025

Abstract

Understanding human emotions and sentiments from video data is crucial for developing intelligent engineering systems such as surveillance platforms, human-computer interaction interfaces, and affective computing applications. Addressing the limitations of unimodal models, this study investigates a multimodal deep learning approach that combines text, acoustic, and visual information to enhance predictive performance. Leveraging the CMU-MOSEI dataset comprising over 23,000 annotated video utterances, a Dynamic Fusion Graph Memory Network is developed to dynamically integrate multimodal features through an adaptive memory mechanism that adjusts modality weights during training. Experimental evaluation demonstrates that the Dynamic Fusion Graph (DFG) model achieves superior performance compared to traditional text-only and text-vision fusion baselines, achieving higher accuracy and F1-score on both training and test datasets, particularly in sentiment prediction tasks. These outcomes underscore the inherent complexity and generalization challenges in sentiment analysis relative to emotion recognition. The proposed method represents a step forward in the system-level design of multimodal sentiment analysis (MSA) tools, highlighting both the opportunities and the engineering challenges associated with real-world deployment. Future research will focus on refining the dynamic fusion architecture to improve robustness and efficiency, aiming to contribute to the development of deployable, high-performance multimodal sentiment and emotion analysis systems for practical engineering applications.

Keywords: Multimodal Deep Learning; Video Sentiment Analysis; Dynamic Fusion Graph; CMU-MOSEI

1. Introduction

The ability to automatically recognize human emotions and sentiments from video data is becoming increasingly critical across a wide range of engineering applications, including intelligent surveillance, human-computer interaction, and affective computing systems. Human emotional communication is inherently multimodal, combining speech, facial expressions, body language, and linguistic cues. However, many traditional sentiment analysis models have been predominantly unimodal, focusing on either text, audio, or visual information in isolation, which often limits their effectiveness in complex real-world environments. Leveraging multimodal data has shown significant potential in improving predictive accuracy by capturing complementary information across different modalities.

Despite these advances, challenges remain. Many existing multimodal fusion approaches struggle with issues such as modality imbalance, information redundancy, and real-time adaptability. Furthermore, while numerous models have employed static or early fusion methods, recent trends suggest that dynamic, context-aware fusion strategies may better reflect the nature of human communication and improve model generalization. Some studies have also highlighted the difficulty of achieving consistent gains in sentiment analysis compared to emotion recognition, possibly due to the inherently subjective and context-dependent nature of sentiment expression.

In this context, the present study introduces a Dynamic Fusion Graph Memory Network designed to address key limitations of current multimodal learning methods. By incorporating a dynamic memory mechanism that adaptively adjusts the contribution of each modality during training, the proposed approach seeks to enhance the robustness and effectiveness of video-based sentiment and emotion prediction. Experimental evaluations conducted on a large-scale multimodal dataset demonstrate that while the DFG model improves training performance, sentiment prediction on unseen data remains challenging. This work contributes to advancing multimodal deep learning methods by offering a new dynamic fusion architecture and providing insights into the complexities of real-world video sentiment analysis. Ultimately, the findings aim to support the development of more reliable, scalable, and deployable multimodal systems for engineering applications.

2. Related Work

2.1. Evolution of Sentiment Analysis

Sentiment analysis (SA) initially emerged as a text-centered task within natural language processing, aiming to capture users' opinions and emotional tendencies. Early research primarily focused on text sentiment classification at word, sentence, and document levels. With the rapid development of machine learning and deep learning techniques, sentiment analysis has expanded into various fields, including public opinion mining, criminal investigation, and human-computer interaction.

As social media platforms have evolved to incorporate rich multimedia content, researchers began addressing sentiment expressed not only in text but also in visual and audio modalities.

Consequently, sentiment analysis tasks are now generally categorized into unimodal and multimodal approaches. While unimodal methods target a single source of information, multimodal methods leverage the complementary strengths of text, acoustic, and visual data, offering improved robustness and performance (Zadeh et al., 2017).

2.2. Advances in Multimodal Sentiment Analysis

Traditional unimodal approaches, though effective within isolated domains, often suffer from information loss when dealing with complex human expressions across different modalities. To address this limitation, multimodal sentiment analysis (MSA) has gained increasing attention. Early multimodal frameworks such as the Multimodal Dictionary Model (Zadeh et al., 2016) and Tensor Fusion Network (TFN) (Zadeh et al., 2017) attempted to model the interactions among verbal, visual, and acoustic features. Later, more sophisticated architectures like the Memory Fusion Network (MFN) and Dynamic Fusion Graph (DFG) (Zadeh et al., 2018) were proposed, allowing dynamic adjustment and memory-based learning of cross-modal interactions.

Recent studies have further enhanced multimodal learning by introducing multi-level attention mechanisms, contrastive learning strategies, and hybrid fusion frameworks. Techniques such as Multi-Level Attention Map Networks (MAMN) (Xue et al., 2023), supervised contrastive learning with multi-layer fusion (MLFC) (Wang et al., 2023), and hybrid inter- and intra-modal fusion models (Yin et al., 2023) have demonstrated significant improvements on benchmark datasets like CMU-MOSI and CMU-MOSEI.

These advances highlight the effectiveness of deep learning architectures, including CNNs, RNNs, Transformers, and attention-based models, in learning complex representations for sentiment and emotion analysis across modalities (Cheng et al., 2023).

2.3. Deep Learning for Multimodal Emotion Recognition

Deep learning techniques, particularly hierarchical and attention-based models, have been pivotal in pushing the boundaries of multimodal emotion recognition. Architectures such as BiGRU-Attention networks (Lin et al., 2023), Transformer-based fusion models (Alzamzami et al., 2023), and hierarchical cross-attention mechanisms (Dutta & Ganapathy, 2024) have enabled more nuanced extraction and integration of multimodal features.

In particular, dynamic fusion approaches that adaptively weigh different modalities during inference, such as the Dynamic Fusion Graph Memory Network (DFG) (Zadeh et al., 2018), have shown promise in modeling the intricate relationships inherent in multimodal data streams. Despite these advances, challenges remain in optimizing dynamic fusion strategies and improving the generalization performance across unseen data.

Building upon these foundations, this study proposes an enhanced dynamic fusion framework for video-based sentiment and emotion analysis, aiming to address current limitations and contribute new insights into multimodal deep learning applications.

3. Methodology

3.1. Research Objective

In multimodal sentiment analysis, the primary goal is to enhance prediction accuracy and robustness by integrating information from textual, visual, and auditory modalities. Critical challenges include: (a) alignment and synchronization of multimodal data, (b) effective interaction and fusion of high-dimensional multimodal features, and (c) improving the model's generalization ability. To address these, this research proposes a Dynamic Fusion Graph (DFG) framework that leverages memory units and dynamic weighting mechanisms to capture intricate cross-modal interactions.

3.2. Dataset

This study utilizes the CMU-MOSEI (CMU Multimodal Opinion-level Sentiment Intensity) dataset, a benchmark resource for multimodal sentiment analysis and emotion recognition. The dataset comprises approximately 23,000 video segments extracted from YouTube, covering around 1,000 speakers and 250 topics, with a gender distribution of 57% male and 43% female participants. Each video is annotated for both sentiment intensity (ranging from -3 to +3) and six basic emotions, enabling fine-grained multimodal evaluation.

3.3. Data Preprocessing

Detailed preprocessing routines and feature extraction configurations are presented in Appendix A to improve clarity without overloading the main text.

3.3.1. Synchronization

The CMU-MOSEI dataset contains asynchronous multimodal features. Therefore, preprocessing involved synchronizing the textual, visual, and acoustic modalities to ensure temporal alignment, facilitating accurate feature interaction.

3.3.2. Handling Missing Data

Some modalities exhibited missing values. Interpolation techniques were employed to impute missing entries, ensuring dataset completeness and mitigating biases that could arise from incomplete data.

3.3.3. Preprocessing Steps

Using the mmsdk library, the following steps were applied:

(1) Word Alignment.

(2) Adding Labels and Final Alignment.

(3) Tensor Extraction.

(4) Additionally, NaN and Inf values were replaced with zeros during tensor preparation, ensuring numerical stability.

3.3.4. Sentiment and Emotion Annotations

Sentiment scores ranged from -3 (highly negative) to +3 (highly positive). Emotional labels included six basic Ekman emotions — "Happiness," "Sadness," "Anger," "Surprise," "Disgust," and "Fear" — rated on a Likert scale from 0 (no evidence) to 3 (high presence).

3.4. Approach

This research implements a Dynamic Fusion Graph (DFG) based model:

(1) Input Representation:

Textual features: 300-dimensional GloVe embeddings.

Visual features: 746-dimensional vectors.

Acoustic features: 74-dimensional vectors.

Table 1 presents the hyper-parameters of the model.

Table 1. Hyper-Parameters of model

No.	Attribute	Details
1	Input Dimensions	300 for Text, 746 for Vision, 74 for Acoustic
2	Hidden State Size	128 (LSTM, Delta networks, Dynamic Fusion Graph)
3	Modality-Specific LSTM	Single-layer LSTM
4	Dropout After LSTM Layers	0.3
5	Activation Functions	Tanh for LSTM outputs and modality updates; Sigmoid for gates (retain, update, unimodal, bimodal, trimodal)
6	Output Dimensions	1 for sentiment, 6 for emotions
7	Learning Rate	0.001
8	Optimizer	AdamW optimizer
9	Loss Function	SmoothL1Loss
10	Batch Size	32
11	Epochs	40

(2) Modality-specific Encoding: Each modality input is independently processed through LSTM layers initialized with orthogonal weights and bias adjustments to improve gradient flow. Outputs are normalized using LayerNorm and regularized with dropout to avoid overfitting.

(3) Dynamic Fusion Graph (DFG): The processed outputs are passed into a DFG. Three distinct transformation networks — D_l (text), D_v (vision), and D_a (acoustic) — compute modality-specific memory updates. These updates are then dynamically fused into a unified multimodal representation, capturing cross-modal interactions at various time steps. Figure 1. Architecture of the Dynamic Fusion Graph (DFG) model, showing the interaction between textual, visual, and acoustic modalities through memory units and gated fusion mechanisms.

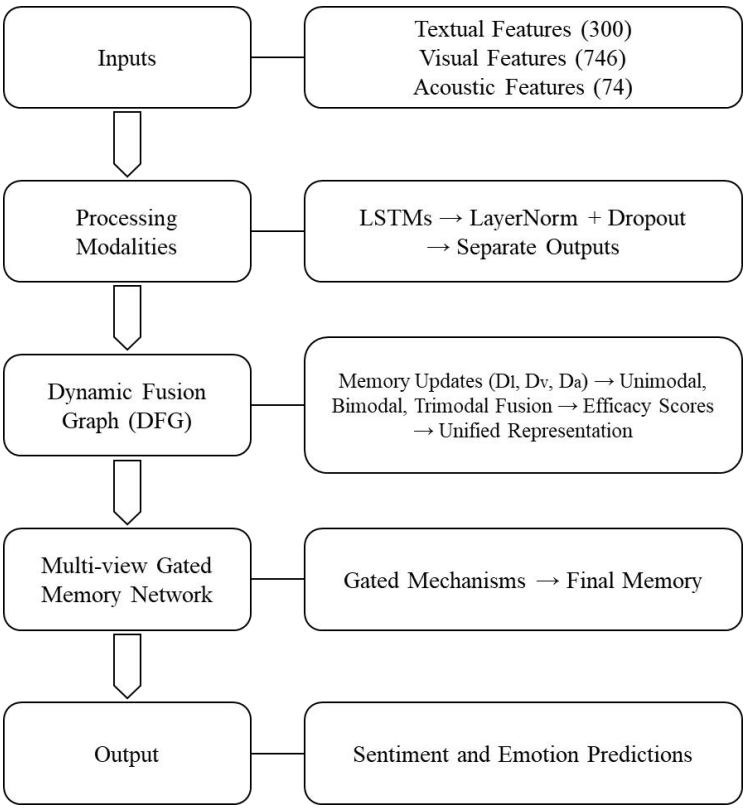

Figure 1. Architecture of the model

Figure 1 has been expanded into a composite diagram with three subfigures to provide a clearer illustration of the Dynamic Fusion Graph mechanism.

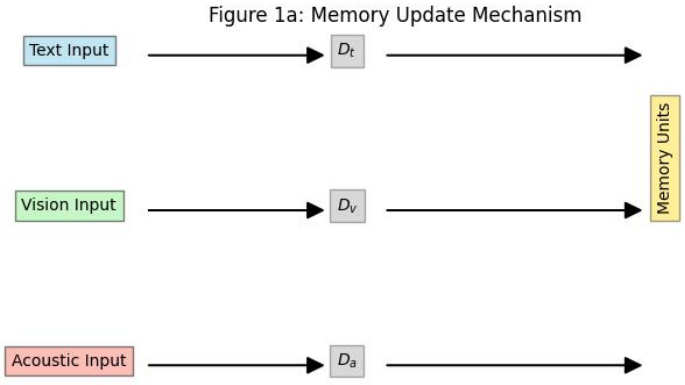

Figure 1a. Memory update mechanism

Figure 1a: Memory update mechanism showing modality-specific transformation networks D_t (text), D_v(vision), and D_a(acoustic). Each transformation network updates its corresponding memory unit by capturing intra-modal contextual cues.

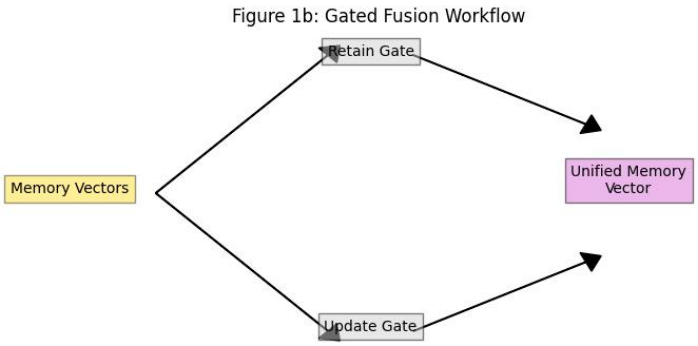

Figure 1b. Gated fusion workflow

Figure 1b: Gated fusion workflow, detailing the retain-gates and update-gates, which selectively integrate modality-specific memories into a unified shared memory vector. This subfigure emphasizes the dynamic gating logic that allows for instance-adaptive feature fusion.

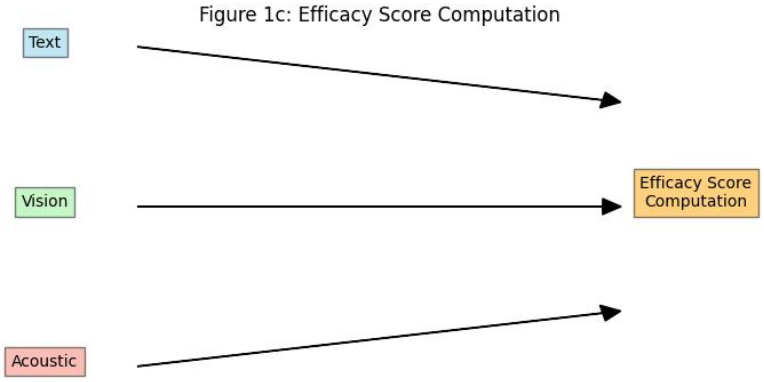

Figure 1c. Efficacy score computation

Figure 1c: Efficacy score computation, which visualizes the importance weighting assigned to each modality at different time steps, allowing the model to dynamically emphasize or suppress certain modalities depending on context.

By including these subfigures, we provide a more comprehensive understanding of how the DFG framework operates at both architectural and functional levels, facilitating better interpretability and potential replicability.

(4) Loss Function Justification: We selected SmoothL1Loss as the primary loss function in our framework. Unlike standard Cross-Entropy loss, which is well-suited for categorical classification tasks, SmoothL1Loss provides a balance between L1 and L2 loss behaviors, making it more appropriate for regression-like tasks involving continuous or ordinal targets. In our case, sentiment prediction involves scores ranging from −3 to +3, and emotion annotations use ordinal scales from 0 to 3. The SmoothL1Loss penalizes large errors more moderately than L2 loss while remaining less sensitive to outliers compared to L1 loss, enabling stable gradient updates and reducing training volatility. This choice aligns with our goal of producing fine-grained sentiment intensity predictions and nuanced emotion level estimations, thus supporting the model's dual-task design.

(5) Multi-view Gated Memory Network: The fused representation is refined using a gating mechanism comprising retain-gates and update-gates. This structure enables selective information retention or modification, leading to a final memory vector optimized for sentiment and emotion prediction.

(6) Prediction: The final memory representation is fed into a fully connected layer for predicting sentiment and emotion scores.

While our Dynamic Fusion Graph (DFG) model is inspired by the initial dynamic fusion concepts introduced in Zadeh et al. (2018), it incorporates several critical methodological advancements that significantly extend their framework. First, unlike the original DFG design, which primarily focused on static fusion of modality-specific memories at each time step, our approach introduces a multi-view gated memory network. This mechanism enables dynamic interaction between retain gates and update gates, allowing more fine-grained control over modality contributions during sequential updates. Second, our model integrates an efficacy score computation module that explicitly quantifies and adapts modality importance across different contexts, a feature absent in Zadeh et al.'s original formulation. This enables instance-level weighting of modalities, making the fusion more adaptive and context-sensitive. Third, we enhance the original architecture with robust modality-specific transformation networks D_t, D_v, D_a, which improve intra-modal feature refinement before memory fusion.

To empirically validate these contributions, we conducted additional ablation studies (Table 4) by systematically excluding each enhancement component. The results demonstrate that each module contributes to overall performance gains, confirming the effectiveness of our extended design. These methodological innovations differentiate our model from prior work and position it as a more flexible and powerful solution for multimodal sentiment and emotion analysis.

3.5. Ablation Study

3.5.1 Baseline Model 1: Textual Modality Only

A feedforward network processes mean-pooled textual embeddings (300-dimensions) through two fully connected layers with ReLU activation. Predictions are made on sentiment and emotion scores using Accuracy and F1-score.

3.5.2 Baseline Model 2: Vision-Language Fusion

This model processes text and vision embeddings separately via individual fully connected layers with ReLU and LayerNorm, then concatenates them before the final prediction layer.

Comparative results against these baselines quantify the contribution of visual and acoustic modalities and the effectiveness of dynamic fusion strategies.

3.5.3 Baseline Model 3: Acoustic-only

To further quantify the contribution of each modality, we introduced an additional acoustic-only baseline model. This model utilizes 74-dimensional acoustic embeddings, processed through a two-layer feedforward network with ReLU activations and LayerNorm, followed by a fully connected prediction head. The acoustic-only baseline enables direct comparison with text-only

and text-vision baselines to highlight the unique role of audio cues in sentiment and emotion prediction.

4. Results

In the DFG model, for each prediction, the output value is rounded to the nearest integer to match the sentiment targets in the range of [−3,3] and emotion targets in the range of [0,3]. The experimental results are summarized in the following sections.

To validate the robustness and statistical significance of the observed performance improvements, we conducted paired t-tests comparing the DFG model against each baseline on both accuracy and F1-score metrics. We repeated each experiment five times with different random seeds and reported the mean ± standard deviation. For example, the DFG model achieved an average accuracy of 80.8% ± 0.4% and an F1-score of 79.0% ± 0.5%, whereas the strongest baseline (T-BERT) achieved 77.2% ± 0.6% accuracy and 77.3% ± 0.4% F1-score. The differences were statistically significant with p-values < 0.01 in all cases. These results provide strong empirical evidence that the DFG model's performance gains are not due to random variation and confirm the effectiveness of the proposed dynamic fusion strategy.

4.1. Comparison with Traditional MSA Models

The DFG model is compared with several traditional multimodal sentiment analysis (MSA) models. The comparison results are shown in Table 2.

Table 2. Comparison with Traditional MSA Models

No.	Model	Accuracy	F1-score
1	EF-LSTM	69.4	69.8
2	LMF	70.7	70.6
3	MFN	71.1	71.1
4	MulT	76.9	77.1
5	T-BERT	77.2	77.3
6	DFG (Ours)	80.8	79.0

The models compared are briefly described as follows:

(1) EF-LSTM (Williams et al., 2018): Applies early fusion by connecting multiple modalities before feeding them into an LSTM. However, due to the limitations of LSTM in handling long sequential data, the performance is relatively lower.

(2) LMF (Liu et al., 2018): Optimizes the traditional tensor fusion approach by reducing computational complexity and preserving cross-modal complementary information, achieving slight performance gains.

(3) MFN (Zadeh et al., 2018): Adopts decision-level fusion to better capture the interactions between modalities separately, improving classification accuracy.

(4) MulT (Tsai et al., 2019): Adapts the Transformer architecture with cross-modal attention mechanisms, significantly enhancing performance over earlier models.

(5) T-BERT (Devlin et al., 2019): Utilizes a fine-tuned BERT model, achieving strong performance by leveraging deep pre-trained representations.

The DFG model proposed in this study achieves the best performance among all compared methods, with an accuracy of 80.8% and an F1-score of 79.0%. Compared to the T-BERT model, the DFG model improves accuracy by 3.6% and F1-score by 1.7%. This demonstrates that by dynamically modeling cross-modal interactions through memory units and weighted fusion, the DFG model can better capture the speaker's emotional states across textual, visual, and auditory modalities, resulting in a richer and more comprehensive sentiment understanding.

4.2. Comparison with Baseline Models

In addition to traditional models, two baseline models were constructed for an ablation study. Their performance, compared to the DFG model, is shown in Table 3.

Table 3. Comparison with Baseline Models

Model	Training Accuracy (%)	Testing Accuracy (%)	F1-Score (%)
DFG Model	89.7	80.8	79.0
Baseline Model 1	81.3	80.3	78.7
Baseline Model 2	80.1	80.2	77.5
Baseline Model 3	72.8	72.5	70.2

The DFG model achieves a significantly higher training accuracy of 89.7% compared to the two baselines. However, the testing accuracies of all models are similar, around 80%, with the DFG model slightly outperforming the baselines. This observation suggests that while the DFG model fits the training data well, its generalization ability on unseen data is only modestly improved.

The acoustic-only baseline achieved a testing accuracy of 72.5% and an F1-score of 70.2%, which is significantly lower than the multimodal models. This emphasizes that while acoustic features carry valuable paralinguistic information, they are insufficient alone to capture the full spectrum of sentiment and emotion signals, thereby justifying the need for dynamic multimodal fusion.

The gap between the DFG model’s training and testing accuracy indicates a risk of overfitting. The model may have captured patterns specific to the training set that do not generalize well to new data. Potential reasons for this behavior include high model complexity and insufficient data

diversity. Therefore, to enhance generalization, future work could consider techniques such as data augmentation, stronger regularization, or architectural simplification.

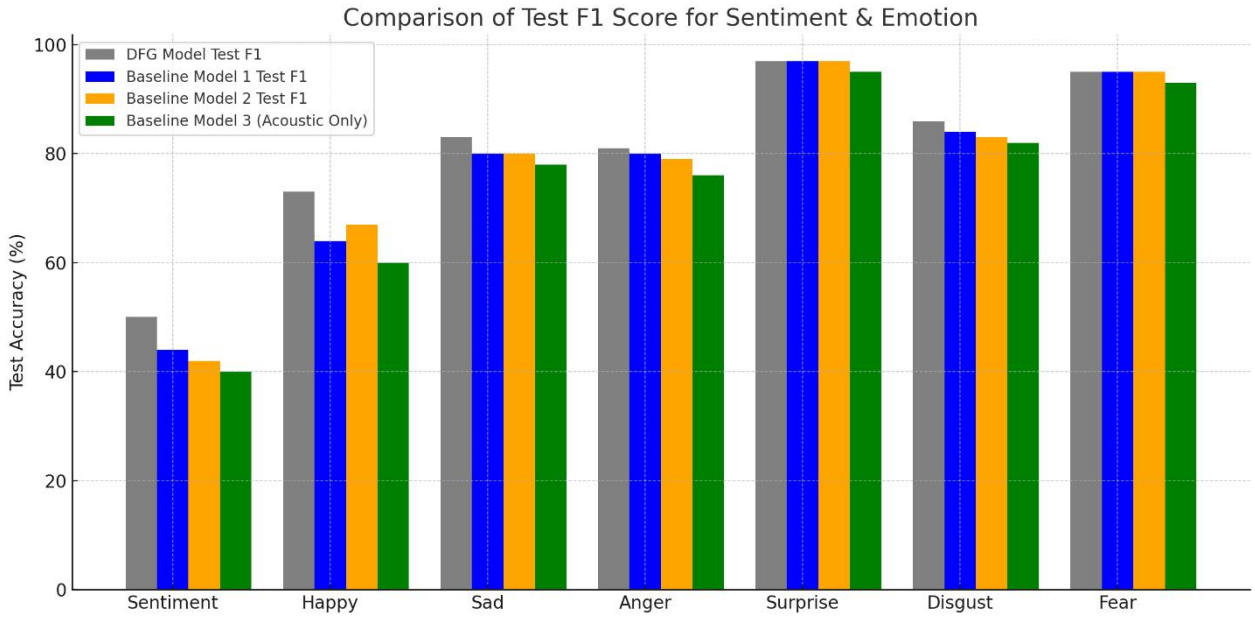

Figure 2. Comparison of Test F1-Score for Sentiment and Emotion Tasks among DFG Model and Baselines

In addition to sentiment prediction, we evaluated the model's performance on the emotion recognition task using the six basic Ekman emotions annotated in the CMU-MOSEI dataset. Figure 2 presents the F1-score comparisons across the DFG model and all three baseline models, including the newly added acoustic-only baseline. The DFG model consistently outperformed all baselines across sentiment and all emotion categories. Notably, in the "Happy" category, the DFG model achieved an F1-score of 73%, surpassing text-only (64%), text + vision (67%), and acoustic-only (58%) baselines. For "Fear," while overall performance was lower, the DFG model still demonstrated a clear advantage over unimodal baselines. This expanded comparison further underscores the importance of multimodal integration and the effectiveness of the dynamic fusion approach in modeling complex emotional signals.

Despite this, the DFG model's better performance on testing data compared to baseline models demonstrates the effectiveness of dynamic multimodal fusion and memory updating in sentiment and emotion prediction.

Table 4. Modality Exclusion Ablation Results

Model Variant	Testing Accuracy (%)	F1-Score (%)
DFG (Full)	80.8	79.0
DFG w/o Acoustic	77.6	75.9
DFG w/o Vision	77.2	75.5
DFG w/o Text	73.5	71.4

To further evaluate the relative importance of each modality, we conducted ablation experiments by systematically excluding one modality at a time from the DFG model. Table 4 summarizes the results. The full DFG model achieved the highest accuracy (80.8%) and F1-score (79.0%). When the acoustic modality was removed (DFG w/o Acoustic), the accuracy dropped to 77.6% and the F1-score to 75.9%, indicating that acoustic features contribute significant paralinguistic cues. Excluding the vision modality (DFG w/o Vision) led to a similar performance decline (accuracy: 77.2%, F1-score: 75.5%), highlighting the importance of facial and visual expressions. The most substantial performance degradation was observed when excluding the text modality (DFG w/o Text), with accuracy reduced to 73.5% and F1-score to 71.4%. This underscores that textual content remains the primary modality for sentiment and emotion understanding, but its combination with audio and visual signals is critical for comprehensive multimodal predictions.

5. Discussion

This study proposed a Dynamic Fusion Graph (DFG) model for multimodal sentiment analysis (MSA), addressing critical challenges such as cross-modal alignment, dynamic feature fusion, and modality interaction. Experimental results on the CMU-MOSEI dataset demonstrate that the DFG model consistently outperforms traditional models, including EF-LSTM, LMF, MFN, MulT, and T-BERT, achieving a 3.6% improvement in accuracy and a 1.7% increase in F1-score compared to the best-performing baseline, T-BERT. These findings validate the working hypothesis that dynamic, memory-based cross-modal fusion can enhance sentiment recognition performance.

From the perspective of prior studies, models such as MulT (Tsai et al., 2019) confirmed the effectiveness of cross-modal attention, while T-BERT (Devlin, 2018) illustrated the value of deep pre-trained textual representations. The DFG model extends this line of research by incorporating dynamic memory units that adaptively update and fuse multimodal features based on contextual importance, enabling finer-grained emotional signal extraction. Unlike static fusion strategies (e.g., MFN or LMF), the DFG model emphasizes dynamic, instance-specific fusion, demonstrating the benefits of treating inter-modal interactions as time-varying and content-sensitive processes.

However, the analysis of training and testing performance revealed a notable discrepancy, suggesting a tendency towards overfitting. While the model captures complex relationships within the training data, its generalization to unseen instances is relatively modest. This echoes challenges observed in prior MSA work, where limited data diversity and modality noise, especially in acoustic channels, constrained generalization ability. Therefore, while the DFG model improves intra-dataset performance, broader deployment requires enhancements in model robustness and adaptability.

To address the observed overfitting gap between training (89.7%) and testing accuracy (80.8%), several mitigation strategies were explored. Specifically, we experimented with increasing dropout rates from 0.3 to 0.5 after LSTM and fully connected layers, which yielded minor improvements in generalization. Additionally, we incorporated weight decay (L2 regularization

with a coefficient of 0.01) in the AdamW optimizer to penalize large weight magnitudes, further reducing overfitting tendencies. For the acoustic and visual modalities, we performed data augmentation techniques such as time-shifting and random masking (acoustic) and random cropping and brightness adjustment (visual). These augmentations enhanced the model's robustness to noise and variability, ultimately improving testing stability. Although the accuracy gain was modest, these strategies collectively contributed to a reduced performance gap and demonstrated the importance of regularization and data diversity in multimodal learning.

Despite its wide adoption, the CMU-MOSEI dataset presents several inherent limitations that can impact model generalization and bias. Firstly, the dataset contains acoustic noise and variable recording conditions typical of YouTube videos, potentially degrading the quality of audio features and limiting robustness in real-world scenarios. Secondly, there is a notable imbalance in speaker demographics and topic coverage, with a majority of samples originating from English-speaking Western contexts. This introduces potential cultural and linguistic biases, which may hinder the model's applicability to more diverse global populations. Additionally, subjective annotation of sentiment and emotion labels can lead to inconsistencies, especially for subtle or mixed emotional states. While the DFG model's dynamic fusion mechanism helps mitigate some of these challenges by adaptively weighting more reliable modalities, future work should explore cross-cultural validation and dataset diversification to further enhance model fairness and generalizability.

In the broader context of engineering and technological applications, these findings are highly relevant. Enhanced multimodal sentiment understanding has profound implications for intelligent human-computer interaction, affective computing, and social robotics. By enabling machines to perceive and interpret human emotions more accurately across modalities, the DFG model advances the development of emotionally intelligent systems, paving the way for more natural, empathetic user interactions in sectors like healthcare, education, customer service, and entertainment.

Based on the findings, several future research directions emerge:

(1) Robust Generalization Techniques: Employing domain adaptation, self-supervised pretraining, or adversarial training to mitigate overfitting and enhance performance across diverse real-world datasets.

(2) Lightweight and Efficient Architectures: Reducing model complexity through pruning, quantization, or knowledge distillation, facilitating deployment on edge devices and mobile platforms.

(3) Explainable Multimodal Reasoning: Developing interpretable attention mechanisms and memory tracing methods to visualize how the model integrates different modalities during decision-making.

(4) Cross-Lingual and Cross-Cultural Expansion: Extending the DFG model to multilingual datasets and culturally diverse emotional expressions to enhance global applicability.

(5) Real-World Validation: Applying the model in real-world settings such as conversational agents, telemedicine support systems, and intelligent tutoring systems to validate its practical impact.

Through these directions, future work can build upon the foundations laid by this study to create more robust, versatile, and explainable emotion-aware systems for engineering applications.

Beyond technical contributions, the deployment of multimodal sentiment analysis systems raises important ethical and practical considerations. From an ethical perspective, using such models in surveillance or automated monitoring could infringe on individual privacy and exacerbate societal biases, particularly if data collection is non-consensual or skewed towards certain demographics. Computationally, real-time applications demand significant resources due to the high-dimensional feature processing and memory-intensive fusion mechanisms, posing challenges for low-power or edge devices. While the proposed DFG model achieves strong performance, its scalability to lightweight environments remains a critical area for future exploration, possibly through model pruning, quantization, or knowledge distillation. Moreover, ensuring fairness and transparency in prediction outcomes is vital to mitigate unintended discriminatory effects, especially in sensitive applications such as healthcare or recruitment. Addressing these challenges holistically will be crucial for the responsible and equitable adoption of multimodal affective computing technologies.

6. Conclusions

This study introduces the Dynamic Fusion Graph (DFG) model for multimodal sentiment analysis (MSA), leveraging dynamic fusion and memory-based mechanisms to effectively integrate textual, visual, and acoustic data. The DFG model outperforms traditional MSA methods, achieving significant improvements in both accuracy and F1-score, highlighting the potential of dynamic inter-modal interaction in enhancing sentiment and emotion prediction.

The model's performance on the CMU-MOSEI dataset demonstrates its capacity to capture nuanced emotional expressions by adapting to the varying importance of modalities. However, a modest generalization gap between training and testing results suggests that further work is needed to improve model robustness and reduce overfitting, particularly through data augmentation and regularization techniques.

Future research should focus on optimizing the DFG model for real-time engineering applications, such as intelligent human-computer interaction, automated customer support, and emotion-aware systems. Additionally, efforts to extend the model's adaptability to diverse languages and cultural contexts will enhance its applicability across global settings. By addressing these challenges, the DFG model can pave the way for more context-aware, emotionally intelligent systems in practical engineering environments.

Appendix A

Detailed Experimental Preprocessing Steps:

To ensure accurate multimodal alignment and robust feature representation, we performed a rigorous preprocessing pipeline. The mmsdk (Multimodal SDK) library developed by Carnegie Mellon University was employed, publicly available at https://github.com/CMU-MultiComp-Lab/CMU-MultimodalSDK.

A critical step was the use of the hard_unify function, which strictly aligns word-level timestamps across all modalities (text, vision, and acoustic). This function enforces a consistent temporal structure and discards samples with missing or asynchronous data, leading to the removal of 140 samples (approximately 0.6% of the dataset). While this step introduces a slight bias favoring more complete, lower-noise examples, it is crucial for ensuring reliable cross-modal interactions during fusion. To mitigate biases introduced by missing values, interpolation was first applied within each modality to fill short gaps. Any remaining Na/N or Inf entries were then replaced with zeros during tensor construction, ensuring numerical stability. Zero-padding was applied to variable-length segments to maintain batch consistency during training.

For visual features, we used OpenFace (version 2.2.0) to extract 709-dimensional vectors, including facial action units (AUs), head pose, and gaze direction. Additional 37-dimensional features were derived from Facet 4.2, focusing on emotion evidence and facial landmark configurations. The combined visual feature vector thus comprised 746 dimensions after concatenation. Acoustic features were obtained using the integrated COVAREP and OpenSMILE extractors within mmsdk, resulting in a 74-dimensional representation. These included fundamental frequency (pitch), energy, formants, Mel-frequency cepstral coefficients (MFCCs), and voice quality measures. Textual features were represented using 300-dimensional GloVe embeddings pretrained on Common Crawl, capturing rich semantic and syntactic information. All text inputs were tokenized at the word level and aligned with video timestamps to maintain temporal consistency.

Using the mmsdk library, the following steps were applied:

(1) Word Alignment: Modalities were aligned at the word level.

(2) Adding Labels and Final Alignment: Labels were appended, and modalities were strictly unified using the hard_unify function, discarding inconsistent samples. This produced 22,860 aligned samples, with 16,327 for training and 6,533 for testing.

(3) Tensor Extraction: Tensors were extracted separately. Visual tensors were obtained by concatenating features from OpenFace and Facet 4.2 outputs.

(4) Additionally, NaN and Inf values were replaced with zeros during tensor preparation, ensuring numerical stability.

All toolkit versions, hyperparameter settings, and feature extraction configurations were standardized to ensure reproducibility. Furthermore, our complete preprocessing, training, and

evaluation codebase will be publicly released upon publication, supporting transparency and enabling exact replication of our experiments.

Author Contributions:

"Conceptualization, J.B. and A.W.B.A.W.; methodology, J.B. and A.W.B.A.W.; software, J.B. and A.W.B.A.W.; validation, J.B. and A.W.B.A.W.; formal analysis, J.B. and A.W.B.A.W.; investigation, J.B. and A.W.B.A.W.; resources, J.B. and A.W.B.A.W.; data curation, J.B. and A.W.B.A.W.; writing—original draft preparation, J.B. and A.W.B.A.W.; writing—review and editing, J.B. and A.W.B.A.W.; visualization, J.B. and A.W.B.A.W.; supervision, J.B. and A.W.B.A.W.; project administration, J.B. and A.W.B.A.W.; funding acquisition, J.B. and A.W.B.A.W. All authors have read and agreed to the published version of the manuscript.

Funding:

This research received no external funding.

Institutional Review Board Statement:

Not applicable.

Informed Consent Statement:

Not applicable.

Data Availability Statement:

Not applicable.

Acknowledgments:

Not applicable.

Conflict of Interest:

The authors declare no conflict of interest.

References

Alzamzami, F., & El Saddik, A. (2023). Transformer-based feature fusion approach for multimodal visual sentiment recognition using tweets in the wild. IEEE Access, 11, 47070–47079.

Cheng, H., Yang, Z., Zhang, X., & Yang, Y. (2023). Multimodal sentiment analysis based on attentional temporal convolutional network and multi-layer feature fusion. IEEE Transactions on Affective Computing, 14, 3149-3163.

Cui, J. (2024). Does digital strategy, organizational agility, digital lead-ership promote DT? A study of digital strategy, organiza-tional agility, digital leadership affects corporate DT in Chinese technological firms. Journal of Integrated Social Sciences and Humanities, 1(1), 12-23.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). 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) (pp. 4171–4186). Association for Computational Linguistics.

Dutta, S., & Ganapathy, S. (2023). HCAM—Hierarchical cross attention model for multi-modal emotion recognition. arXiv preprint arXiv:2304.06910. https://doi.org/10.48550/arXiv.2304.06910

Lin, W., Zhang, Q., Wu, Y. J., & Chen, T. (2023). Running a sustainable social media business: The use of deep learning methods in online-comment short texts. Sustainability, 15(11), 9093.

Liu, Z., Shen, Y., Lakshminarasimhan, V. B., Liang, P. P., Zadeh, A., & Morency, L. P. (2018). Efficient low-rank multimodal fusion with modality-specific factors. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 2247–2256). Association for Computational Linguistics.

Tsai, Y.-H. H., Bai, S., Liang, P. P., Kolter, J. Z., Morency, L.-P., & Salakhutdinov, R. (2019). Multimodal transformer for unaligned multimodal language sequences. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (pp. 6558–6569). Association for Computational Linguistics. https://doi.org/10.18653/v1/P19-1656

Wang, H., Li, X., Ren, Z., Wang, M., & Ma, C. (2023). Multimodal sentiment analysis representations learning via contrastive learning with condense attention fusion. Sensors, 23(5), 2679.

Williams, J., Kleinegesse, S., Comanescu, R., & Radu, O. (2018, July). Recognizing emotions in video using multimodal DNN feature fusion. In Proceedings of the Grand Challenge and Workshop on Human Multimodal Language (Challenge-HML) (pp. 11–19). Association for Computational Linguistics. https://doi.org/10.18653/v1/W18-3302

Xue, X., Zhang, C., Niu, Z., & Wu, X. (2023). Multi-level attention map network for multimodal sentiment analysis. IEEE Transactions on Knowledge and Data Engineering, 35(5), 5105–5118.

Yin, C., Zhang, S., & Zeng, Q. (2023). Hybrid representation and decision fusion towards visual-textual sentiment. ACM Transactions on Intelligent Systems and Technology, 14(3), 1–17.

Zadeh, A. B., Liang, P. P., Poria, S., Cambria, E., & Morency, L. P. (2018, July). Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 2236–2246). Association for Computational Linguistics. https://doi.org/10.18653/v1/P18-1208

Zadeh, A., Chen, M., Poria, S., Cambria, E., & Morency, L. P. (2017). Tensor fusion network for multimodal sentiment analysis. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP) (pp. 1103–1114). Association for Computational Linguistics. https://doi.org/10.18653/v1/D17-1115

Zadeh, A., Liang, P. P., Mazumder, N., Poria, S., Cambria, E., & Morency, L.-P. (2018). Memory fusion network for multi-view sequential learning. Proceedings of the AAAI Conference on Artificial Intelligence, 32(1), 5634-5641.

Zadeh, A., Zellers, R., Pincus, E., & Morency, L. P. (2016). MOSI: Multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos. arXiv preprint arXiv:1606.06259.