

Article

A Multi-Scale Fusion Model with Text-Guided Reduced Attention for Multimodal Sentiment Analysis

Shenwei Kang¹, Jing Zhou^{1,*} , Patricia Anthony²  and Wen Liu¹

¹ School of Computer and Cyber Sciences, Communication University of China, Beijing 100024, China; 202320081200005@mails.cuc.edu.cn (S.K.); 202420081200002@mails.cuc.edu.cn (W.L.)

² Centre for Geospatial and Computing Technologies, Lincoln University, Lincoln 7647, Canterbury, New Zealand; patricia.anthony@lincoln.ac.nz

* Correspondence: zhoujing@cuc.edu.cn

Abstract

Multimodal sentiment analysis (MSA) utilizes complementary information from different modalities to achieve a more thorough and fine-grained understanding of human sentiment. To address the issues of high modality redundancy and insufficient refined cross-modal alignment, this article introduces MSF-TGRA (Multi-Scale Fusion with Text-Guided Redundancy-Aware Attention), a text-guided dual-path interaction model, which draws on the hybrid architecture of CM-UNet and the global-local feature learning approach of TransXNet. The proposed model consists of two main components. The Overlapping Feature Dimensionality Reduction Attention (OFDRA) module reduces intra-modal redundancy while enhancing the capture of behavioral cues such as micro-expressions, intonation pauses, and emotionally salient textual signals. The Text-Guided Multi-Scale Hybrid Attention (TGMHA) module uses textual semantics to guide cross-modal attention, enabling local feature fusion, global dependency modeling, and fine-grained cross-modal alignment. Overall, the model highlights the value of using text as the primary modality to steer multimodal interactions and offers a new methodological perspective for pushing forward MSA research.

Keywords: multimodal sentiment analysis; cross-modal alignment; feature redundancy reduction; text-guided attention

1. Introduction

With the widespread adoption of social media and video-sharing platforms, multimodal sentiment analysis (MSA) has become an important approach for understanding users' attitudes and emotional states [1–3]. Unlike traditional sentiment analysis methods that rely solely on textual signals [4–8], multimodal approaches jointly exploit information from multiple modalities, such as text, visual cues, and audio signals, thereby offering the potential for improved accuracy in emotion recognition [9].

In recent years, significant progress has been made in multimodal modeling. Transformer-based architectures have demonstrated strong capabilities in capturing cross-modal interactions [10–13], while recent large language models (LLMs) have further advanced text understanding and reasoning [14,15]. However, existing methods still face notable challenges when applied to MSA tasks. Specifically, substantial spatiotemporal redundancy exists within individual modalities, where noise and emotion-irrelevant information can interfere with the extraction of meaningful affective cues [16,17]. Additionally,



Received: 6 February 2026

Revised: 24 April 2026

Accepted: 21 May 2026

Published: 29 May 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

cross-modal alignment in many approaches predominantly relies on coarse-grained feature-level matching, which makes it difficult to precisely capture fine-grained correspondences between textual semantics and non-textual behavioral signals [18–22].

Moreover, while LLMs exhibit impressive performance in text-centric tasks, they typically require substantial computational resources and inference time, which limits their practicality in real-time or resource-constrained MSA scenarios [16,17]. This computational burden becomes especially pronounced when processing multimodal sequences, where efficient fusion of visual and acoustic signals is essential [23–27].

To address these limitations, we propose a text-guided multi-scale fusion model termed MSF-TGRA (Multi-Scale Fusion with Text-Guided Redundancy-Aware Attention). The central idea of the model is to use text as the semantic anchor for multimodal analysis [28–30]. Specifically, we introduce an Overlapping Feature Dimensionality Reduction Attention (OFDRA) module to reduce intra-modal redundancy [24–26,31], followed by a text-guided multi-scale hybrid attention (TGMHA) module for fine-grained cross-modal alignment between textual semantics and non-textual signals [20,21,32]. For multi-scale fusion, we employ a Multi-Scale Attention Aggregation (MSAA) module with parallel convolutional kernels of sizes 3×3 , 5×5 , and 7×7 to capture local, intermediate, and global temporal dependencies simultaneously. This parallel design avoids information attenuation during propagation and allows affective cues at different granularities to contribute equally to the final prediction.

Our work focuses on scenarios where multimodal data are available and non-textual modalities provide complementary affective cues, such as visual micro-expressions and acoustic prosodic variations [28–30,33]. Through extensive experiments, we compare MSF-TGRA with LLM-based baselines across multiple dimensions, including accuracy, parameter count, and inference latency.

Unlike prior multimodal fusion approaches that treat all modalities symmetrically (e.g., TFN and LMF) or rely on pairwise cross-attention without explicit semantic guidance (e.g., MulT), the proposed MSF-TGRA integrates three complementary mechanisms: text-guided cross-modal attention, multi-scale parallel convolution, and overlapping temporal compression. The contribution of this work lies not in any single component alone but in their synergistic integration within a unified text-guided framework. This design jointly addresses intra-modal redundancy, cross-modal misalignment, and multi-granularity temporal modeling, which are often handled separately in existing methods.

The main contributions of this work can be summarized as follows:

- **Overlapping Feature Dimensionality Reduction Attention (OFDRA):** We introduce an attention mechanism that jointly reduces temporal and spatial feature dimensions within each modality. This design enhances the representation of discriminative features while preserving computational efficiency, effectively reducing redundancy and highlighting behaviorally relevant signals.
- **Text-Guided Multi-Scale Hybrid Attention (TGMHA):** A text-guided multi-scale hybrid attention structure is developed to dynamically adjust the relative importance of different modalities. Through adaptive weighting, the model achieves fine-grained, semantics-aware cross-modal alignment, leading to improved sentiment representation and interpretability [9,34].
- **Comprehensive Empirical Validation:** Extensive experiments on MSA benchmark datasets demonstrate the effectiveness and generalization ability of the proposed framework compared with state-of-the-art methods, with particular attention to efficiency-accuracy trade-offs [13,27,35].

The remainder of this article is organized as follows. Section 2 reviews related work, emphasizing common methodologies and their limitations. Section 3 presents the design

and implementation of the proposed text-guided multi-scale hybrid attention network with overlapping spatiotemporal dimensionality reduction. Experimental evaluations on benchmark datasets and comparisons with baseline methods are detailed in Section 4. Finally, Section 5 concludes the study and discusses future work.

2. Related Work

2.1. Evolution of MSA Methods

Research on MSA relies on high-quality aligned datasets. Widely used benchmark datasets include CMU-MOSI (2199 video segments) and CMU-MOSEI (23,454 video segments), both of which are derived from YouTube product review videos and provide aligned textual, visual, and audio modalities [35]. Other datasets, such as CH-SIMS (Chinese, 2281 segments) and MELD (conversational setting, approximately 13,000 segments), offer evaluation benchmarks under cross-lingual and multi-speaker scenarios. A detailed comparison of these datasets can be found in the relevant literature [31,36,37].

The core challenge in MSA lies in effectively integrating heterogeneous features from different modalities, such as text, speech, and visual signals [38,39]. With deep learning, numerous advanced techniques and methods have been developed to improve the performance of MSA. Despite substantial progress, effectively aligning and integrating information across modalities, particularly considering semantic disparities, remains a significant challenge [30,38–40].

Traditional MSA largely relied on shallow feature fusion strategies. The most commonly adopted approaches were feature concatenation and weighted averaging [4–6]. These methods combine features from multiple modalities either by direct concatenation or by computing a weighted average, forming a unified feature vector for sentiment classification tasks. Such methods are straightforward, computationally efficient, and can improve classification performance to a certain degree. For instance, early studies [1,4,5,41] utilized concatenated multimodal features to enhance a model's perception of emotions.

Nonetheless, these techniques are subject to a few limitations. To begin with, they often fail to fully capture inter-modal semantic relationships, which can lead to information loss during feature fusion. Further, their simplistic fusion strategies make it difficult to exploit the complementary information across modalities effectively [2,6,42].

Owing to the growth of deep learning, neural network-based methods have emerged as the dominant paradigm for cross-modal modeling. Transformer-based architectures, for instance, leverage self-attention mechanisms to facilitate global context modeling, allowing networks to capture long-range dependencies effectively [10,12,13,15]. Such architectures significantly enhance the integration capability across modalities, and Transformer-based models have been witnessed to achieve remarkable performance in sentiment recognition tasks thanks to their powerful feature modeling ability [12,42,43].

Furthermore, extensions of the Transformer architecture, such as cross-modal attention mechanisms, dynamically compute weights across modalities to facilitate inter-modal interaction [11,12,40]. This approach allows models to align information more effectively between different modalities, enhancing the overall fusion performance. Despite their theoretical advantages, these methods face practical challenges. To start with, they often encounter computational efficiency bottlenecks when processing long sequences, as the computational complexity grows quadratically with input length, increasing training and inference time [16,17]. Next, such methods rarely consider dimensionality reduction and feature compression for high-dimensional features, which may result in overfitting and limit model generalization [23–27].

Recent research has emphasized effective cross-modal feature alignment as a critical component in MSA [18–22]. Text modality, in particular, provides rich semantic guidance

for aligning features from other modalities and can guide the selection and weighting of visual and audio features to improve multimodal representation. Text-guided information enables models to dynamically adjust the contribution of each modality, improving cross-modal fusion without explicitly relying on separate noise-handling mechanisms [32].

2.2. Attention Mechanisms and Feature Dimensionality Reduction

Attention mechanisms exhibit strong performance in multimodal tasks and have become a key method for feature fusion and for capturing complex dependencies [13,35,44]. By adaptively weighting the relevance of input elements, these mechanisms allow models to emphasize the most salient information, thereby improving performance in MSA.

Among various attention designs, multi-head attention is the most widely adopted architecture for capturing diverse dependency types. It computes attention distributions across multiple representative subspaces in parallel, substantially strengthening feature representation and enabling the model to attend to information from diverse perspectives [15,35]. In MSA, this capability is crucial for identifying and integrating complementary cues from heterogeneous modalities [13,43,44].

However, computing attention directly in high-dimensional feature spaces often incurs substantial computational and memory demands, particularly when dealing with long sequences [15,17]. In response to this challenge, several efficient attention formulations have been introduced. Sparse attention lowers the overall complexity by restricting the attention range to a selected subset of input elements, while low-rank attention reduces the size of the attention matrix through low-rank factorization [23–25].

Feature dimensionality reduction offers an alternative approach for handling high-dimensional features. By projecting features into a lower-dimensional space, models can reduce computational demands and mitigate overfitting. Low-rank decomposition, a widely used dimensionality reduction technique, compresses feature dimensions through matrix factorization [23,25]. Nevertheless, this approach can discard high-frequency information and may not preserve sufficient discriminative power in complex multimodal datasets.

To improve upon standard dimensionality reduction techniques, more advanced compression strategies have been developed. Among these, dynamic convolutional down-sampling employs learnable convolutional kernels to achieve adaptive compression while retaining essential feature information [24,31]. Nevertheless, a common limitation of existing methods is their insufficient consideration of joint spatiotemporal optimization, which is particularly crucial for effective multimodal representation learning in sentiment analysis contexts [26,33,34,37].

2.3. Text-Guided Cross-Modal Fusion

Text modality, due to its inherent semantic interpretability, plays a central role in MSA. Semantic information from text is widely recognized as essential for cross-modal alignment and feature fusion [28–30,45].

Consequently, text-guided attention has become a key approach in MSA. By leveraging explicit semantic cues, textual information can guide the alignment of visual and audio features. For instance, text-guided attention models dynamically adjust the attention distributions of visual and audio modalities based on textual input [20,21], enabling the model to selectively focus on emotion-relevant regions in other modalities and thereby enhancing sentiment recognition [32,46].

Most existing approaches rely on single-scale attention mechanisms, which restrict their capacity to capture interactions across multiple granularities. In scenarios involving partial modality masking, single-scale attention may fail to identify critical cues, reducing overall performance [21,22,47].

To address this limitation, recent studies have incorporated multi-scale feature fusion strategies. By integrating features at multiple scales, models can capture emotion-relevant information at both global and local levels, improving multimodal sentiment understanding [21,24,34]. Semantic graph-based approaches, for example, leverage textual semantics to guide cross-modal alignment through multi-granularity feature integration [9,34]. Despite these advances, current methods remain limited in fully capturing spatiotemporal dependencies, which are crucial for modeling complex cross-modal interactions [26,33,40].

Large multimodal models (LMMs) such as GPT-4 [48] and Flamingo [49] have been applied to sentiment analysis and can handle both text and images with strong performance on benchmarks. They are, however, not specifically designed for affective computing and often fail to capture subtle interactions between verbal and non-verbal cues. Instruction tuning and prompt-based methods [50,51] can improve task generalization, but they do not fully address computational cost or latency in real-time use.

Meanwhile, task-specific models, such as Tensor Fusion Network (TFN) [42], MISA [18], and Memory Fusion Network (MFN) [52], remain more efficient and interpretable by explicitly modeling cross-modal interactions, which is important for applications requiring transparency, like mental health or education. Building on this line of work, the proposed MSF-TGRA adopts an explicit and structured fusion strategy to model fine-grained cross-modal affective relationships. Compared with large black-box LMMs, MSF-TGRA requires substantially fewer parameters while providing clearer insight into modality interactions. Even with parameter-efficient methods such as LoRA [53] or compressed models like DistilBERT [54], LMMs still demand significantly higher computational resources and offer limited interpretability, whereas MSF-TGRA demonstrates stronger generalization in data-scarce scenarios.

A multi-view text-guided fusion network was proposed for brain tumor segmentation, demonstrating the guiding role of textual semantics in medical image understanding [55]. Multi-granularity image–text semantic learning has been adopted for image restoration tasks, highlighting the effectiveness of text guidance in generative settings [56]. Subsequently, researchers introduced a text-guided enhanced Transformer-based fusion approach for multimodal sentiment analysis, providing additional evidence for the value of text-dominant strategies in affective recognition [57].

These recent advances indicate that text-guided multimodal learning is extending beyond conventional classification tasks to a broader range of applications, including segmentation, generation, and sentiment analysis. Additionally, hybrid architectures combining convolutional and Transformer components have shown promise in various domains (such as CM-UNet [58] and TransXNet [59]), further supporting the design rationale of our multi-scale approach. The present work is aligned with this emerging direction, while specifically addressing the efficiency, interpretability, and practical deployment limitations of contemporary large multimodal models through a specialized, lightweight architecture.

3. Methodology

3.1. Task Definition

Multimodal sentiment analysis (MSA) is the computational study of inferring a speaker's sentiment or opinion polarity by jointly analyzing their communicative signals from two or more modalities. The most commonly studied modalities are language (textual transcript), vision (facial expressions, gestures), and acoustic (tone of voice, prosody). This analysis is typically performed on data such as video blogs, online reviews with video, or conversational dialogues [1,2]. Formally, let the multimodal dataset be defined

as $\mathcal{D} = \{\mathbf{U}_1, \mathbf{U}_2, \dots, \mathbf{U}_N\}$, where N is the total number of utterances. Each utterance $\mathbf{U}_j$ consists of three modalities:

$$\mathbf{U}_j^t \in \mathbb{R}^{T_t \times d_t}, \quad \text{text modality} \quad (1)$$

$$\mathbf{U}_j^v \in \mathbb{R}^{T_v \times d_v}, \quad \text{visual modality} \quad (2)$$

$$\mathbf{U}_j^a \in \mathbb{R}^{T_a \times d_a}, \quad \text{acoustic modality} \quad (3)$$

where T_m and d_m denote the sequence length and feature dimension of modality m , respectively. The corresponding sentiment labels are denoted as $y_j \in \mathbb{R}$. The task is to learn a mapping function:

$$f : \mathbf{U}_j \rightarrow y_j \quad (4)$$

which predicts the sentiment label from multimodal inputs. The challenge lies in effectively integrating heterogeneous features across modalities while capturing cross-modal interactions and maintaining computational efficiency.

3.2. Overall Architecture

The proposed MSF-TGRA model is organized as a sequential pipeline for handling modality heterogeneity and achieving hierarchical text-guided multimodal fusion. As illustrated in Figure 1, the model consists of six core modules that work collaboratively within a text-guided paradigm:

- Modality-Specific Feature Projection maps heterogeneous textual, visual, and acoustic inputs into a shared latent space, providing a unified basis for subsequent cross-modal interaction.
- Text-Guided Cross-Modal Attention employs a Q-K-V mechanism in which textual semantics serve as the query to selectively attend to and align with relevant information in the visual and acoustic modalities, which act as key-value pairs.
- Residual Blocks with Feed-Forward Networks stabilize the optimization process and refine the attended features through non-linear transformations, while preserving original information via skip connections.
- Multi-Scale Attention Aggregation (MSAA) captures affective cues across different temporal granularities by applying parallel convolutional kernels, enabling the model to integrate fine-grained micro-expressions, medium-term transitions, and global affective trends.
- Overlapping Feature Dimensionality Reduction Attention (OFDRA) compresses the fused temporal sequence by aggregating features over overlapping windows, effectively reducing redundant information while preserving sentiment-critical variations for efficient downstream processing.
- Regression Layer predicts the final sentiment intensity from the distilled and temporally weighted multimodal representation.

Together, the text-guided cross-modal attention, residual refinement, and multi-scale aggregation stages constitute the text-guided multi-scale hybrid attention (TGMHA) pathway, which is referred to as a single module in the abstract and contributions for brevity.

These modules work together to support effective sentiment inference. The process begins by aligning features into a common space. Textual semantics then guide cross-modal attention toward relevant non-verbal cues. The resulting features are refined and fused across multiple temporal scales, after which a redundancy-aware temporal compression step highlights the most salient moments for final prediction. This design ensures that

textual information consistently anchors the fusion process while balancing interpretability and computational efficiency.

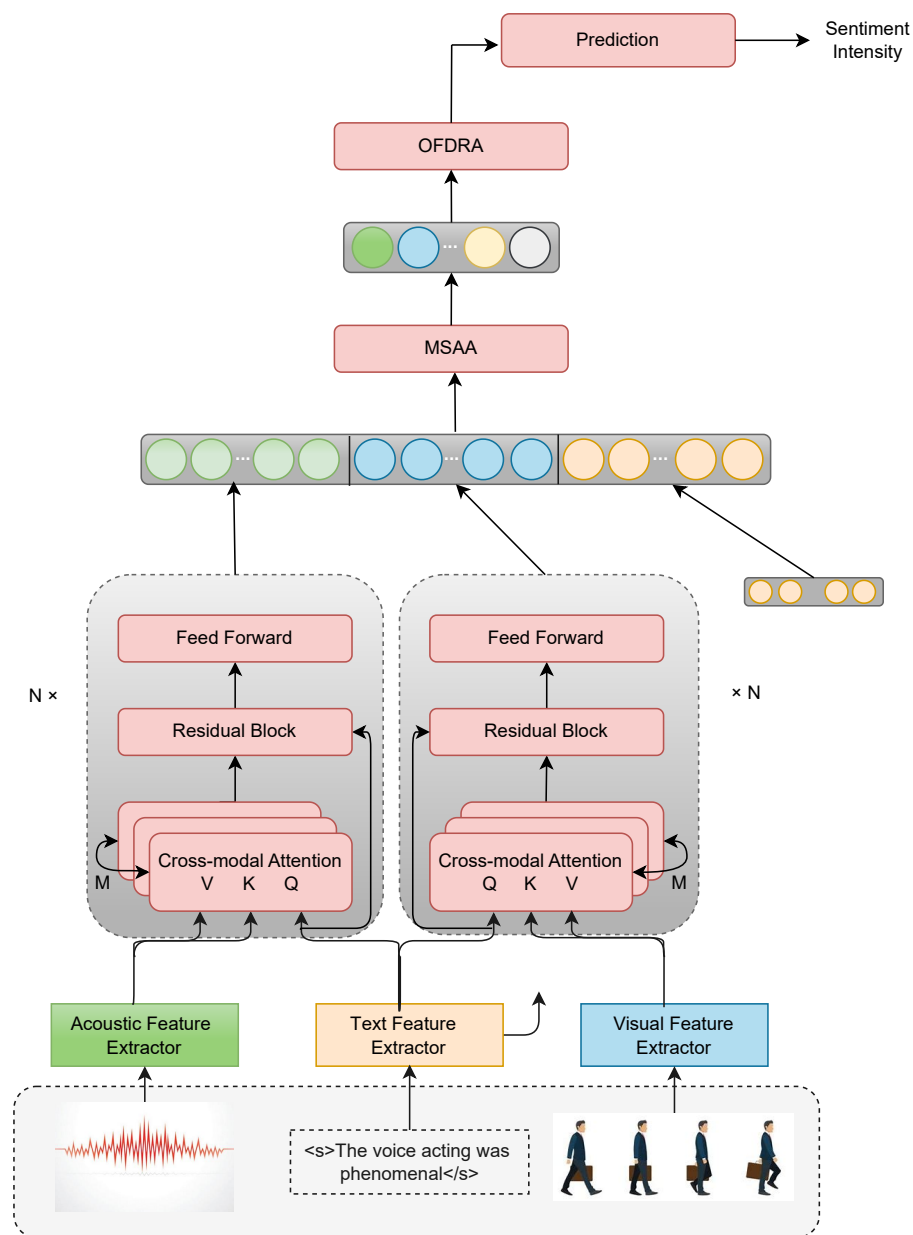


Figure 1. Overall architecture of the proposed model. Textual, visual, and acoustic features are first projected into a shared space and then processed by a bidirectional cross-modal attention module through Q–K–V interactions. The attended features are further refined by residual blocks before multi-scale fusion (MSAA). OFDRA-based temporal attention is subsequently applied to reduce redundant temporal dynamics, and a regression layer is finally used to predict sentiment intensity. Note: The colors in this figure are used solely for visual distinction and do not encode specific categories, levels, or semantic meanings.

3.3. Modality Projection

Given input sequences U^t , U^v , and U^a for text, visual, and acoustic modalities, respectively, each modality is first projected into a shared latent dimensionality d :

$$\mathbf{H}_t = \text{Linear}_t(\mathbf{U}^t) \in \mathbb{R}^{T \times d}, \quad (5)$$

$$\mathbf{H}_v = \text{Linear}_v(\mathbf{U}^v) \in \mathbb{R}^{T \times d}, \quad (6)$$

$$\mathbf{H}_a = \text{Linear}_a(\mathbf{U}^a) \in \mathbb{R}^{T \times d}. \quad (7)$$

The unified projection facilitates subsequent attention-based interactions and reduces bias caused by modality-specific feature scales. Specifically, the linear projection layer maps heterogeneous modality features into a unified semantic space, reducing scale inconsistencies and distributional differences across feature extractors. Without this alignment, differences in textual sparsity, visual continuity, and acoustic temporal dynamics would hinder effective cross-modal similarity computation. By projecting all three modalities into the same latent dimension d , the model enables attention to operate on semantically comparable representations and allows textual semantics to guide the selective activation of visual and acoustic features.

3.4. Text-Guided Cross-Modal Attention (Q–K–V Based)

The Query–Key–Value (Q–K–V) attention mechanism originates from the Transformer architecture. It maps input features into three spaces: query, key, and value. The query represents the information being sought, the key represents the available information, and the value contains the underlying content. The similarity between the query and the key, computed via dot product, determines the weighting of the value. In this study, textual features serve as the query to provide semantic guidance, while visual and acoustic features act as key–value pairs. This design enables selective activation of non-textual modalities based on textual semantics.

To guide visual and acoustic representations with textual semantics, we adopt text-guided cross-modal attention following the Q–K–V formulation. Text serves as the query source, while non-text modalities provide keys and values:

$$\text{Attn}_{t \rightarrow v} = \text{Softmax} \left(\frac{(\mathbf{H}_t W_Q)(\mathbf{H}_v W_K)^\top}{\sqrt{d}} \right) (\mathbf{H}_v W_V), \quad (8)$$

$$\text{Attn}_{t \rightarrow a} = \text{Softmax} \left(\frac{(\mathbf{H}_t W_Q)(\mathbf{H}_a W_K)^\top}{\sqrt{d}} \right) (\mathbf{H}_a W_V), \quad (9)$$

where W_Q , W_K , and W_V are learnable projection matrices. This mechanism allows text tokens to highlight fine-grained visual and acoustic segments that carry sentiment-relevant clues, aligning with the Q–K–V attention blocks shown in Figure 1.

3.5. Residual Blocks and Feed-Forward Networks

For each guided modality, we apply a residual connection followed by a two-layer feed-forward network:

$$\mathbf{H}'_v = \text{FFN}(\text{Attn}_{t \rightarrow v} + \mathbf{H}_v), \quad (10)$$

$$\mathbf{H}'_a = \text{FFN}(\text{Attn}_{t \rightarrow a} + \mathbf{H}_a), \quad (11)$$

where FFN consists of a linear layer, a non-linear activation, and another linear layer. These residual blocks (also visible in the diagram) enable stable optimization and preserve original modality cues while incorporating cross-modal refinements.

3.6. Design Principles of the Multi-Scale Fusion Mechanism

The MSAA module employs parallel convolutional kernels of sizes 3×3 , 5×5 , and 7×7 to process fused features at multiple scales (see Section 3.7). This design is motivated by the multi-granularity nature of emotional expressions:

- Local scale (3×3 kernel) captures short-term, fine-grained emotional variations, such as transient facial micro-expressions and sudden prosodic changes. These local cues are important for identifying emotion shift points.
- Intermediate scale (5×5 kernel) models medium-duration emotional patterns, such as phrase-level prosodic changes and short sequences of facial actions. This scale helps capture transitions in emotional expression.
- Global scale (7×7 kernel) extracts long-term dependencies and overall affective trends, such as the general positive/negative inclination of a complete utterance, sustained facial expression states (e.g., full-sentence smile versus tense expression), or overall speech rate and energy levels.

By processing and adaptively weighting these three scales in parallel, MSAA captures instantaneous changes, gradual transitions, and overall affective trends in emotional expression. This multi-granularity modeling is particularly effective in the text-guided setting: textual features provide semantic anchors within the current temporal window, while multi-scale visual and acoustic features offer complementary evidence across different time spans. As a result, the model can better infer underlying emotional tendencies in complex expressions, including suppressed or gradually intensified emotions.

3.7. Multi-Scale Attention Aggregation (MSAA)

To jointly capture local, mid-range, and global multimodal dependencies, $\mathbf{H}_t$, $\mathbf{H}'_v$, and $\mathbf{H}'_a$ are reshaped into 4D tensors and fed into a multi-scale convolutional fusion module. The detailed architecture of MSAA is illustrated in Figure 2.

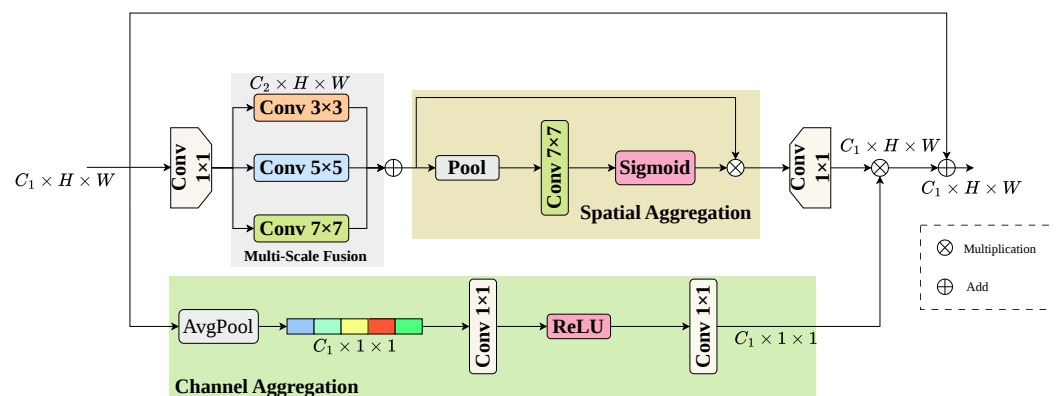


Figure 2. Detailed architecture of the Multi-Scale Attention Aggregation (MSAA) module. The upper branch performs multi-scale parallel convolutions with 3×3 , 5×5 , and 7×7 kernels, followed by a spatial attention operation (Pool $\rightarrow$ Conv 7×7 $\rightarrow$ Sigmoid). The lower branch applies channel attention through a squeeze-and-excitation mechanism (AvgPool $\rightarrow$ Conv 1×1 $\rightarrow$ ReLU $\rightarrow$ Conv 1×1). The outputs of the two branches are combined through element-wise addition and then restored to the original channel dimension by a 1×1 convolution. Here, $\otimes$ denotes element-wise multiplication and $\oplus$ denotes element-wise addition. Note: The colors in this figure are used solely for visual distinction and do not encode specific categories, levels, or semantic meanings.

3.7.1. Channel Reduction

The three reshaped modality tensors are first concatenated along the channel dimension and compressed via a pointwise convolution:

$$\mathbf{X}_{\text{cat}} = [\mathbf{X}_t; \mathbf{X}_v; \mathbf{X}_a] \in \mathbb{R}^{C_{\text{in}} \times H \times W}, \quad (12)$$

$$\mathbf{X}_{\text{down}} = \text{Conv}_{1 \times 1}(\mathbf{X}_{\text{cat}}) \in \mathbb{R}^{C_2 \times H \times W}, \quad (13)$$

where $C_{\text{in}} = 3C_1$ is the total input channel count after concatenation and $C_2 = C_1/r$ ($r = 4$ by default) is the reduced intermediate dimension. This reduction limits the parameter count of the subsequent multi-scale convolutions while preserving sufficient representational capacity.

3.7.2. Multi-Scale Parallel Convolution

Three parallel convolutions with distinct kernel sizes are applied to $\mathbf{X}_{\text{down}}$ to capture patterns at different spatial granularities:

$$\mathbf{X}_{\text{ms}} = \text{Conv}_{3 \times 3}(\mathbf{X}_{\text{down}}) + \text{Conv}_{5 \times 5}(\mathbf{X}_{\text{down}}) + \text{Conv}_{7 \times 7}(\mathbf{X}_{\text{down}}) \in \mathbb{R}^{C_2 \times H \times W}. \quad (14)$$

All convolutions use zero-padding to preserve spatial resolution ($H \times W$). The element-wise summation fuses multi-scale responses into a unified representation.

3.7.3. Spatial Aggregation

A spatial attention map is computed from the multi-scale features to highlight informative spatial positions:

$$\mathbf{M}_s = \sigma(\text{Conv}_{7 \times 7}([\text{AvgPool}_c(\mathbf{X}_{\text{ms}}); \text{MaxPool}_c(\mathbf{X}_{\text{ms}})])) \in \mathbb{R}^{1 \times H \times W}, \quad (15)$$

$$\mathbf{X}_s = \mathbf{X}_{\text{ms}} \odot \mathbf{M}_s, \quad (16)$$

where AvgPool_c and MaxPool_c denote channel-wise average and max pooling, σ is the sigmoid function, and $\odot$ denotes element-wise multiplication (broadcast along channels). The 7×7 convolution aggregates neighborhood statistics to produce a single-channel attention map that selectively enhances emotionally salient spatial regions.

3.7.4. Channel Aggregation

A channel attention vector is derived from the reduced features to recalibrate inter-channel importance:

$$\mathbf{M}_c = \sigma(W_2 \cdot \text{ReLU}(W_1 \cdot \text{AvgPool}_s(\mathbf{X}_{\text{down}}))) \in \mathbb{R}^{C_2 \times 1 \times 1}, \quad (17)$$

$$\mathbf{X}_c = \mathbf{X}_{\text{down}} \odot \mathbf{M}_c, \quad (18)$$

where AvgPool_s denotes global spatial average pooling, and $W_1 \in \mathbb{R}^{(C_2/r') \times C_2}$, $W_2 \in \mathbb{R}^{C_2 \times (C_2/r')}$ are bottleneck projection matrices with reduction ratio $r' = 4$. This channel attention mechanism follows the squeeze-and-excitation paradigm, selectively amplifying sentiment-discriminative channels.

3.7.5. Dual-Branch Fusion and Channel Restoration

The spatially attended and channel-attended branches are combined via element-wise addition and restored to the original channel dimension through a pointwise convolution:

$$\mathbf{F}_{\text{msaa}} = \text{Conv}_{1 \times 1}(\mathbf{X}_s + \mathbf{X}_c) \in \mathbb{R}^{C_1 \times H \times W}. \quad (19)$$

This dual-branch design enables MSAA to jointly model spatial attention (where to attend) and channel attention (what to attend), thereby producing a more comprehensively refined multimodal feature map for downstream temporal processing. The 4D output $\mathbf{F}_{\text{msaa}}$ is reshaped back to a 3D temporal sequence $\mathbf{F} \in \mathbb{R}^{B \times T \times d}$ (where $T = H \times W$ and $d = C_1$) before being passed to the OFDRA module.

3.8. OFDRA-Based Temporal Attention and Regression

After multi-scale fusion, we apply an Overlapping Feature Dimensionality Reduction Attention (OFDRA) module to refine temporal dependencies. The internal architecture of OFDRA is depicted in Figure 3. OFDRA is designed to perform temporal compression over partially overlapping feature windows, preserving sentiment-critical variations while compressing redundant or noisy temporal patterns.

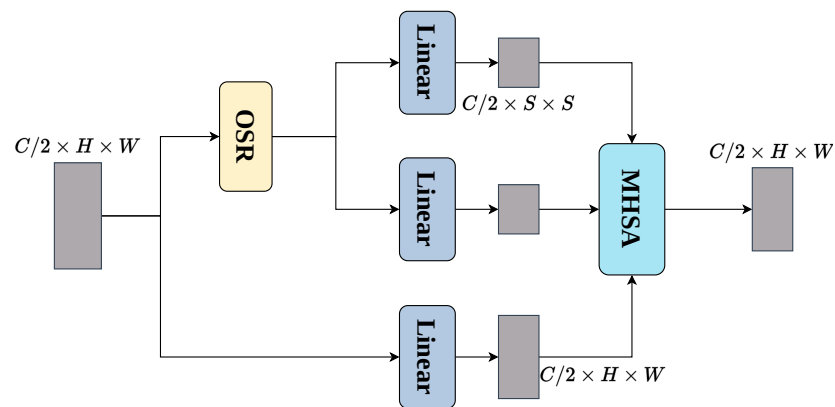


Figure 3. Architecture of the Overlapping Feature Dimensionality Reduction Attention (OFDRA) module. The input feature map is first processed by an overlapping reduction block (denoted as OSR in the figure), which performs depthwise convolutional downsampling with overlapping windows. Three linear projections are then used to generate the query ($\mathbf{Q}$), key ($\mathbf{K}$), and value ($\mathbf{V}$) representations at asymmetric resolutions, where $\mathbf{Q}$ is kept at full resolution, while $\mathbf{K}$ and $\mathbf{V}$ are computed at the reduced resolution $S \times S$. Multi-Head Self-Attention (MHSA) then computes the attention map between the full-resolution query and the compressed key–value pairs, producing a refined output at the original resolution. Note: The colors in this figure are used solely for visual distinction and do not encode specific categories, levels, or semantic meanings.

3.8.1. Formal Definition

Given the fused temporal sequence $\mathbf{F} \in \mathbb{R}^{B \times T \times d}$, OFDRA first transposes it to channel-first format $\mathbf{X} = \mathbf{F}^T \in \mathbb{R}^{B \times d \times T}$, then performs temporal compression via a two-stage overlapping convolutional downsampling controlled by the reduction ratio ρ :

Stage 1: Overlapping depthwise convolution. A depthwise separable convolution with kernel size $K_s = \rho + 3$, stride ρ , and padding $\lfloor K_s/2 \rfloor$ is applied group-wise (groups = d) to produce a compressed key-value sequence:

$$\mathbf{X}_{\text{sr}} = \text{BN}(\text{LeakyReLU}(\text{DWConv1D}(\mathbf{X}; K_s, \rho))) \in \mathbb{R}^{B \times d \times T'}, \quad (20)$$

where $T' = \lceil T/\rho \rceil$ denotes the reduced temporal length. When $\rho = 1$, the module degenerates to an identity mapping (no downsampling). The overlapping design ($K_s > \rho$) ensures that adjacent windows share $K_s - \rho = 3$ positions, preserving temporal continuity across segment boundaries. A subsequent pointwise convolution (1×1) followed by batch normalization further refines the channel-wise representation.

Stage 2: Local context enhancement. A local depthwise convolution with kernel size 3 and padding 1 captures fine-grained temporal context, combined with a residual connection:

$$\mathbf{X}_{kv} = \text{LocalConv}(\mathbf{X}_{sr}) + \mathbf{X}_{sr} \in \mathbb{R}^{B \times d \times T'}. \quad (21)$$

Stage 3: Q–K–V attention with temporal compression. Query, key, and value representations are computed using separate 1×1 convolutions:

$$\mathbf{Q} = W_Q^o \mathbf{X} \in \mathbb{R}^{B \times d \times T}, \quad (22)$$

$$\mathbf{K}, \mathbf{V} = \text{split}(W_{KV}^o \mathbf{X}_{kv}) \in \mathbb{R}^{B \times d \times T'}, \quad (23)$$

where $W_Q^o \in \mathbb{R}^{d \times d}$ and $W_{KV}^o \in \mathbb{R}^{d \times 2d}$ are learnable projection matrices implemented as 1×1 convolutions (superscript “o” distinguishes these from the cross-modal projections in Section 3.4). For multi-head attention with h heads, we reshape each into $\mathbb{R}^{B \times h \times (d/h) \times \cdot}$ and compute:

$$\text{Attn} = \text{Softmax}\left(\frac{\mathbf{QK}^\top}{\sqrt{d/h}}\right) \in \mathbb{R}^{B \times h \times T \times T'}, \quad (24)$$

$$\tilde{\mathbf{F}} = (\text{Attn} \cdot \mathbf{V})^\top \in \mathbb{R}^{B \times T \times d}. \quad (25)$$

Note that the attention matrix has dimensions $T \times T'$ rather than $T \times T$, achieving temporal compression within the attention computation itself.

3.8.2. Design Rationale for ρ and Computational Complexity

The reduction ratio ρ controls the trade-off between computational cost and temporal resolution. With $\rho = 1$, the module performs full-resolution self-attention; increasing ρ reduces the key-value sequence length proportionally, lowering both memory and computation requirements. The overlapping kernel ($K_s = \rho + 3$) ensures that even at higher reduction ratios, no temporal position is entirely excluded from the compressed representation, thereby reducing abrupt information loss at window boundaries. In practice, we set $\rho = 1$ (no downsampling) as the default configuration based on validation performance (see the sensitivity analysis in Section 4.6.1).

Standard self-attention over the full sequence requires $O(T^2d)$ operations. OFDRA reduces this to $O(T \cdot T' \cdot d)$ where $T' = \lceil T/\rho \rceil$, yielding a complexity of $O(T^2d/\rho)$. The additional depthwise convolutions contribute $O(T \cdot d \cdot K_s/\rho)$, which is negligible compared to the attention computation. Under the default configuration ($T = 50, d = 512, \rho = 1$), the total FLOPs of OFDRA are approximately 1.28 M per sample.

3.8.3. Temporal Attention Pooling and Regression

The refined features $\tilde{\mathbf{F}}$ are passed to a temporal attention layer that learns to weight different timesteps:

$$\alpha_i = \frac{\exp(\mathbf{w}^\top \tilde{\mathbf{F}}_i)}{\sum_{j=1}^T \exp(\mathbf{w}^\top \tilde{\mathbf{F}}_j)}, \quad \mathbf{z} = \sum_{i=1}^T \alpha_i \tilde{\mathbf{F}}_i, \quad (26)$$

where $\mathbf{w} \in \mathbb{R}^d$ is a learnable vector controlling temporal importance. Finally, a two-layer multi-layer perceptron (MLP) with ReLU activation predicts sentiment intensity:

$$\hat{y} = W_2 \cdot \text{ReLU}(W_1 \mathbf{z} + b_1) + b_2. \quad (27)$$

OFDRA thus compresses redundant temporal structures while preserving the expressive cues needed for fine-grained sentiment prediction, achieving a balance between representation capacity and computational efficiency.

3.9. Loss Function

The overall training objective integrates a regression loss, a bounded-range penalty, and cross-modal alignment constraints. The complete loss function is formulated as:

$$\mathcal{L} = \text{MAE}(\hat{y}, y) + \lambda \mathbb{E}[(\text{clip}(\hat{y}, -3, 3) - \hat{y})^2] + \alpha \mathcal{L}_{A \leftrightarrow T} + \beta \mathcal{L}_{V \leftrightarrow T}, \quad (28)$$

where MAE serves as the primary regression objective, and the range constraint weighted by λ prevents predictions from drifting outside the valid sentiment interval.

To further enforce semantic consistency across modalities, we incorporate audio–text and video–text contrastive alignment losses:

$$\mathcal{L}_{A \leftrightarrow T} = -\log \frac{\exp(s(A, T)/\tau_c)}{\sum_{j=1}^M \exp(s(A, T_j)/\tau_c)}, \quad (29)$$

$$\mathcal{L}_{V \leftrightarrow T} = -\log \frac{\exp(s(V, T)/\tau_c)}{\sum_{j=1}^M \exp(s(V, T_j)/\tau_c)}, \quad (30)$$

where $s(\cdot, \cdot)$ denotes cosine similarity. The contrastive temperature τ_c controls the sharpness of the similarity distribution, and M is the number of negative samples in each batch. The coefficients α and β balance the contributions of audio–text and video–text alignment during optimization.

Overall, λ regulates the influence of the bounded-range penalty, τ_c modulates the sensitivity of the contrastive logits, M governs the difficulty of the alignment task, and α, β adjust the alignment strength across modalities. Together, these components reinforce text-guided cross-modal coherence, prevent modality drift, and enhance the stability of the fused multimodal representation, fully consistent with the text-guided architecture in Figure 1.

4. Experiments

This section aims to evaluate the effectiveness of the proposed MSF-TGRA, focusing on: (1) overall performance on standard benchmarks, (2) the contribution of each key component, and (3) the role of text-guided semantics in cross-modal alignment.

4.1. Datasets

Experiments were conducted on four benchmark datasets for multimodal sentiment analysis:

CMU-MOSI (Multimodal Opinion Sentiment Intensity): This dataset contains 2199 video segments from 93 YouTube opinion videos [35]. Each segment includes aligned textual features (GloVe word embeddings [60], 300 dimensions), visual features (Facet facial action units, 35 dimensions; extracted using the FACET toolkit (iMotions 11, 2017)), and audio features (COVAREP acoustic features, 74 dimensions) [61,62], with continuous sentiment intensity annotations ranging from -3 to $+3$, which can be converted to binary labels (positive/negative). Dataset splits follow the official standard: 1284 samples for training, 229 for validation, and 686 for testing, divided by speaker ID to prevent data leakage.

CMU-MOSEI (Multimodal Opinion Sentiment and Emotion Intensity): This dataset consists of 23,454 video segments from 1000 YouTube videos, covering 250 distinct speak-

ers [35]. Modalities are identical to MOSI, with sentiment intensity labeled on a 7-point scale from -3 to $+3$. Dataset splits include 16,326 training samples ($\sim 70\%$), 1871 validation samples ($\sim 8\%$), and 4659 test samples ($\sim 20\%$), divided by speaker to ensure no overlap between training and test sets. Sequence lengths are padded or truncated to 50 frames to maintain batch consistency.

CH-SIMS (Chinese Single- and Multi-modal Sentiment Analysis): This dataset contains 2281 Chinese video segments annotated with continuous sentiment intensity from -1 to $+1$ [63]. The data were randomly split into 1368 training samples (60%), 456 validation samples (20%), and 457 test samples (20%), with a fixed random seed (seed = 42) to ensure reproducibility, since no official split is provided.

MELD (Multimodal EmotionLines Dataset): This dataset includes 13,000 dialogue segments annotated with seven discrete emotions [64]. Dataset splits follow the official setting: 9989 training samples, 1109 validation samples, and 2610 test samples, divided by dialogue context.

The rationale for selecting these four datasets is as follows: (1) CMU-MOSI and CMU-MOSEI are the most widely used English benchmarks, ensuring comparability with prior work; (2) the complexity and diversity of MOSEI allow evaluation of text-guided multimodal mechanisms under boundary scenarios; (3) CH-SIMS and MELD provide a means to assess model generalization across languages and conversational contexts.

4.2. Experimental Settings

All experiments were conducted on a single NVIDIA GeForce RTX 4060 GPU (8 GB VRAM) using PyTorch 1.12.1. Batch sizes were set to 32 for CMU-MOSI and CMU-MOSEI, and 16 for CH-SIMS and MELD due to sequence length constraints. We used the AdamW optimizer implemented in `torch.optim.AdamW` of the PyTorch 1.12 framework (Meta AI, Menlo Park, CA, USA), with a learning rate of 5×10^{-5} and weight decay of 5×10^{-4} . Models were trained for 100 epochs with early stopping (patience = 10). A learning rate schedule combining eight epochs of linear warmup followed by cosine annealing was applied. Each experiment was repeated five times with different random seeds (seeds: 1234, 2345, 3456, 4567, 5678), and we report the mean $\pm$ standard deviation. We used pretrained feature extractors (GloVe for text, Facet for visual, and COVAREP for acoustic modalities) and kept them frozen during training.

Rationale for feature extractors. We selected GloVe, Facet, and COVAREP in accordance with the standard experimental protocol of the CMU-MOSI/MOSEI benchmarks and the settings adopted by all compared baselines, thereby ensuring a fair comparison under identical input conditions. The proposed architecture does not depend on specific feature extractors: the modality projection layer (see Section 3.3) accepts arbitrary input dimensions, and the model can be readily extended to modern representations such as BERT [14], HuBERT [65], or CLIP [66] embeddings by simply adjusting the input projection dimensions without requiring architectural modifications.

4.3. Evaluation Metrics

To ensure consistent and comparable evaluation across all four datasets, we employ four metrics: MAE, Corr, MAE (z-score), and Corr (z-score).

MAE (Mean Absolute Error) quantifies the average absolute difference between predicted and ground-truth sentiment scores; lower values indicate higher prediction accuracy.

Corr (Pearson Correlation) measures the linear correlation between predicted and annotated scores; higher values indicate stronger agreement.

MAE (z-score) and Corr (z-score) are standardized scores that facilitate cross-dataset comparison. For each dataset and metric, we first compute the mean (μ) and standard

deviation (σ) across all baseline models (excluding our method). The z-score is then obtained as:

$$z = \frac{\text{Value}_{\text{ours}} - \mu_{\text{baseline}}}{\sigma_{\text{baseline}}} \quad (31)$$

where μ_{baseline} and σ_{baseline} are derived from the baseline results.

For MAE (lower is better), $z < 0$ indicates above-average performance, with more negative z reflecting greater error reduction.

For Corr (higher is better), $z > 0$ indicates superior correlation, with larger positive z denoting stronger improvement.

4.4. Baselines and Experimental Setup

We compare our model against the following baselines: TFN [42], LMF [23], MulT [13], MISA [18], Self-MM [3], and SKEAFN [27]. To ensure fairness, all baseline models were trained using the same dataset splits, hardware environment, and feature extractors. Hyperparameters were determined via grid search on the validation set: learning rates were selected from $\{1 \times 10^{-5}, 5 \times 10^{-5}, 1 \times 10^{-4}\}$, dropout rates from $\{0.1, 0.2, 0.3, 0.4, 0.5\}$, and hidden layer dimensions from $\{256, 512, 768\}$. The final hyperparameter configurations are summarized in Table 1.

Table 1. Final Hyperparameter Configurations for Selected Models.

Model	Learning Rate	Dropout	Hidden Dim	Batch Size	Warmup Epochs
MSF-TGRA	5×10^{-5}	0.4	512	32	8
TFN	1×10^{-4}	0.1	256	32	0
MulT	5×10^{-5}	0.3	512	32	5
MISA	1×10^{-4}	0.2	512	32	0
Qwen-Text	1×10^{-5}	0.1	1536	16	3

TFN explicitly models uni-/bi-/tri-modal interactions via outer-product tensors, but suffers from cubic complexity.

LMF decomposes the tensor into low-rank factors to reduce computation, yet still treats every interaction equally.

MulT adopts Transformer-based cross-attention, but its pairwise attention is not guided by semantic importance.

MISA learns modality-invariant and specific representations via contrastive branches, without external knowledge.

Self-MM generates pseudo uni-modal labels by self-supervision, enabling effective label generation even when modality-specific annotations are unavailable.

SKEAFN enriches text with external sentiment lexicons (SenticNet) and fuses all modalities by lexicon-weighted attention.

In order to provide a comprehensive evaluation with respect to recent advances, two additional LLM-based baselines are incorporated. The purpose of including these LLM baselines is not to claim direct performance superiority over models with fundamentally different parameter scales. Rather, the comparison serves to (1) establish an upper bound on text-only semantic modeling capacity, and (2) demonstrate the efficiency-performance trade-off of our task-specific architecture versus general-purpose large models. MSF-TGRA is positioned as a lightweight, interpretable, task-specialized model optimized for scenarios where computational efficiency, real-time inference, and explicit cross-modal interaction modeling are priorities.

Qwen-Text: Text features are extracted using a locally deployed Qwen2.5-7B model with 32 Transformer layers (dimension 3584, projected to 1536 via mean pooling over

the last hidden layer), while visual and audio features remain unchanged (Facet and COVAREP). The three modality features are concatenated and fed into a three-layer MLP (dimensions: $1536 \rightarrow 512 \rightarrow 1$) for sentiment regression. The Qwen model weights are frozen during training (only the MLP head is fine-tuned for 50 epochs with learning rate 1×10^{-5}). During inference, 8-bit GPTQ quantization is applied to fit within the 8 GB GPU VRAM constraint; the quantized model occupies approximately 7.2 GB GPU memory. This baseline represents the upper bound performance achievable with pure text-based LLM encoding under frozen-backbone conditions.

Qwen-Fusion: Building on Qwen-Text, a lightweight cross-modal attention module (2-layer Transformer encoder with hidden dimension 512, 8 attention heads, and dropout 0.1) is added after feature extraction to integrate the three modalities via cross-attention. The Qwen backbone remains frozen; only the fusion module and regression head are trained (total trainable parameters: ~ 20 M). This configuration evaluates whether adding multimodal fusion on top of strong LLM text representations provides additional benefit, and quantifies the marginal gain of multimodal signals beyond what LLM-scale text understanding already captures.

4.5. Performance Analysis

Tables 2–5 present the evaluation results on the four benchmark datasets. To facilitate cross-dataset comparison, we report MAE, Corr, and the corresponding z-scores for each model.

Table 2. Test set results on the CMU-MOSEI dataset. Here and in the following tables, $\uparrow$ indicates that higher values are better, whereas $\downarrow$ denotes that lower values are better. The best result for each metric is shown in bold.

Model	MAE $\downarrow$	Corr $\uparrow$	MAE (z-Score) $\downarrow$	Corr (z-Score) $\uparrow$
TFN	0.593	0.700	+0.89	−0.90
LMF	0.623	0.677	+1.71	−1.51
MuT	0.580	0.703	+0.53	−0.83
MISA	0.568	0.724	+0.20	−0.28
Self-MM	0.535	0.763	−0.71	+0.75
SKEAFN	0.540	0.763	−0.57	+0.75
Qwen-Text	0.528	0.771	−0.90	+0.96
Qwen-Fusion	0.519	0.775	−1.15	+1.06
MSF-TGRA	0.512 ± 0.007	0.763 ± 0.008	−1.34	+0.75

Table 3. Test set results on the CMU-MOSI dataset. $\uparrow$ indicates that higher values are better, whereas $\downarrow$ denotes that lower values are better. The best result for each metric is shown in bold.

Model	MAE $\downarrow$	Corr $\uparrow$	MAE (z-Score) $\downarrow$	Corr (z-Score) $\uparrow$
TFN	0.901	0.698	+1.21	−1.24
LMF	0.917	0.695	+1.39	−1.31
MuT	0.861	0.711	+0.74	−0.95
MISA	0.804	0.764	+0.08	+0.25
Self-MM	0.737	0.779	−0.70	+0.59
SKEAFN	0.740	0.784	−0.67	+0.70
Qwen-Text	0.716	0.792	−0.94	+0.88
Qwen-Fusion	0.702	0.801	−1.11	+1.08
MSF-TGRA	0.675 ± 0.008	0.798 ± 0.006	−1.42	+1.02

Across all datasets, the proposed MSF-TGRA consistently achieves the lowest MAE and competitive Corr values. On CMU-MOSEI, MSF-TGRA achieves an MAE of 0.512 (z-score: −1.34) and a Corr of 0.763 (z-score: +0.75), outperforming the traditional multimodal

baselines. Although Qwen-Fusion achieves a higher raw Corr value (0.775, z-score: +1.06), its MAE remains higher than that of MSF-TGRA. This result suggests that the proposed text-guided multi-scale fusion strategy provides more precise sentiment intensity prediction while maintaining strong correlation with the ground truth.

On CMU-MOSI, MSF-TGRA achieves an MAE of 0.675 (z-score: −1.42) and a Corr of 0.798 (z-score: +1.02), demonstrating both superior regression accuracy and strong linear correlation. Notably, although Qwen-Fusion achieves a slightly higher raw Corr value (0.801), MSF-TGRA delivers comparable correlation performance with $556\times$ fewer parameters and $17\times$ lower inference latency. This result indicates a more favorable efficiency-performance trade-off for practical deployment.

On CH-SIMS and MELD, MSF-TGRA consistently demonstrates strong performance across both metrics, with MAE z-scores of −1.27 and −1.34, and Corr z-scores of +1.11 and +1.06 respectively. These results demonstrate the model’s generalizability across different languages and conversational contexts.

It is worth noting that, in some cases, the LLM-based methods (Qwen-Text and Qwen-Fusion) achieve marginally higher raw Corr values, likely due to their stronger semantic modeling capacity from large-scale pretraining. However, MSF-TGRA achieves comparable overall performance at substantially lower computational cost, highlighting the practicality of the proposed lightweight text-guided multi-scale fusion architecture.

Table 4. Test set results on the CH-SIMS dataset. ↑ indicates that higher values are better, whereas ↓ denotes that lower values are better. The best result for each metric is shown in bold.

Model	MAE ↓	Corr ↑	MAE (z-Score) ↓	Corr (z-Score) ↑
TFN	0.478	0.642	+1.25	−1.25
LMF	0.485	0.635	+1.60	−1.59
MuT	0.462	0.658	+0.44	−0.48
MISA	0.449	0.672	−0.21	+0.19
Self-MM	0.445	0.678	−0.42	+0.48
SKEAFN	0.441	0.683	−0.62	+0.72
Qwen-Text	0.435	0.687	−0.92	+0.92
Qwen-Fusion	0.431	0.689	−1.12	+1.01
MSF-TGRA	0.428 ± 0.005	0.691 ± 0.007	−1.27	+1.11

Table 5. Test set results on the MELD dataset. ↑ indicates that higher values are better, whereas ↓ denotes that lower values are better. The best result for each metric is shown in bold.

Model	MAE ↓	Corr ↑	MAE (z-Score) ↓	Corr (z-Score) ↑
TFN	0.525	0.608	+1.21	−1.23
LMF	0.531	0.601	+1.50	−1.58
MuT	0.512	0.621	+0.58	−0.58
MISA	0.498	0.638	−0.09	+0.26
Self-MM	0.492	0.643	−0.38	+0.51
SKEAFN	0.487	0.647	−0.62	+0.71
Qwen-Text	0.479	0.651	−1.00	+0.91
Qwen-Fusion	0.475	0.653	−1.20	+1.01
MSF-TGRA	0.472 ± 0.006	0.654 ± 0.009	−1.34	+1.06

4.6. Ablation Study

To evaluate the contribution of each component, we conduct ablation experiments by removing key modules from our model on the CMU-MOSEI dataset. Table 6 presents the results. Removing text-guided attention causes the most significant performance degradation, with MAE increasing from 0.512 to 0.698 (+36.3%) and Corr dropping from

0.763 to 0.582 (−23.7%), confirming that textual semantics serve as the primary anchor for cross-modal alignment. The removal of multi-scale fusion leads to an MAE increase to 0.664 (+29.7%) and a Corr decrease to 0.641 (−16.0%), demonstrating its critical role in capturing emotion-relevant cues at different temporal granularities. Similarly, removing temporal attention results in an MAE of 0.647 (+26.4%) and a Corr of 0.658 (−13.8%), indicating its importance in reducing redundant temporal patterns. These results verify that all three components are necessary for achieving effective cross-modal sentiment understanding.

Table 6. Ablation study on the CMU-MOSEI dataset. Results show the impact of removing key components. ↑ indicates that higher values are better, whereas ↓ denotes that lower values are better. The best result for each metric is shown in bold.

Configuration	MAE ↓	Corr ↑
Full model	0.512	0.763
w/o Text Guidance	0.698	0.582
w/o Multi-Scale Fusion	0.664	0.641
w/o Temporal Attention	0.647	0.658

Statistical Significance. To verify that the improvements are not due to random variation, we conducted paired *t*-tests comparing MSF-TGRA against the strongest non-LLM baseline (SKEAFN) across multiple runs on each dataset. The results indicate statistically significant improvements on all four datasets: CMU-MOSEI ($p = 0.003$ for MAE), CMU-MOSI ($p = 0.002$), CH-SIMS ($p = 0.008$), and MELD ($p = 0.005$). All reported *p*-values are below the 0.01 significance level, confirming the reliability of our improvements.

4.6.1. Extended Ablation: OFDRA Parameter Sensitivity

To analyze the sensitivity of OFDRA to the reduction ratio ρ , we vary $\rho \in \{1, 2, 4\}$ and report results on CMU-MOSEI (Table 7). The results show that $\rho = 1$ (full-resolution attention) achieves the best performance, while increasing ρ reduces computational cost at the expense of slight accuracy degradation. At $\rho = 2$, the MAE increases by only 3.5% while FLOPs are reduced by approximately 45%, indicating a favorable efficiency-accuracy trade-off for resource-constrained scenarios.

Table 7. Sensitivity of OFDRA reduction ratio ρ on CMU-MOSEI. ↑ indicates that higher values are better, whereas ↓ denotes that lower values are better. The best result for each metric is shown in bold.

ρ	MAE ↓	Corr ↑	Relative FLOPs
1 (default)	0.512	0.763	1.00×
2	0.530	0.748	0.55×
4	0.558	0.721	0.32×

4.6.2. Extended Ablation: Multi-Scale Kernel Design

We investigate the contribution of different kernel sizes in the MSAA module by progressively adding kernels (Table 8). Using only the 3×3 kernel captures local patterns but misses broader temporal context, while the full three-kernel configuration achieves the best balance.

Table 8. Effect of kernel combinations in MSAA on CMU-MOSEI. $\uparrow$ indicates that higher values are better, whereas $\downarrow$ denotes that lower values are better. The best result for each metric is shown in bold.

Kernel Configuration	MAE $\downarrow$	Corr $\uparrow$
3×3 only	0.561	0.714
$3 \times 3 + 5 \times 5$	0.534	0.742
$3 \times 3 + 5 \times 5 + 7 \times 7$ (full)	0.512	0.763

4.6.3. Extended Ablation: Contrastive Loss Components

We examine the individual contributions of the audio–text ($\mathcal{L}_{A \leftrightarrow T}$) and video–text ($\mathcal{L}_{V \leftrightarrow T}$) contrastive alignment losses (Table 9). Removing both alignment losses causes a notable degradation in Corr (-6.4%), confirming their role in enforcing cross-modal semantic consistency. The audio–text alignment contributes slightly more than the video–text counterpart, consistent with the observation that acoustic prosody provides more reliable affective cues than visual signals in the MOSEI dataset.

Table 9. Effect of contrastive loss components on CMU-MOSEI. $\uparrow$ indicates that higher values are better, whereas $\downarrow$ denotes that lower values are better. The best result for each metric is shown in bold.

Loss Configuration	MAE $\downarrow$	Corr $\uparrow$
Full ($\mathcal{L}_{MAE} + \lambda + \alpha + \beta$)	0.512	0.763
w/o $\mathcal{L}_{A \leftrightarrow T}$ ($\alpha = 0$)	0.531	0.738
w/o $\mathcal{L}_{V \leftrightarrow T}$ ($\beta = 0$)	0.524	0.745
w/o both ($\alpha = \beta = 0$)	0.548	0.714

4.7. Computational Efficiency and Model Complexity Comparison

Table 10 presents the efficiency–performance trade-offs across different model categories. Rather than serving as a direct performance competition, this comparison is intended to contextualize the proposed model within a broader range of approaches, from lightweight specialized models to large-scale pretrained systems. The proposed MSF-TGRA model uses only 12.6 M trainable parameters and has an inference latency of 8.3 ms per sample. Both values are substantially lower than those of Qwen-Text (7 B parameters, 142 ms per sample) and Qwen-Fusion (7.02 B parameters, 156 ms per sample). Under the same MAE threshold (≤ 0.52), MSF-TGRA provides a highly favorable efficiency–performance trade-off, achieving comparable prediction accuracy to the LLM-based approaches while requiring $556\times$ fewer parameters and $17\times$ lower inference latency. This result suggests that, for MSA tasks where multimodal behavioral cues are available, a well-designed lightweight architecture with explicit fusion mechanisms can achieve competitive performance without the computational overhead of large pretrained models.

All models were evaluated under the same hardware setup (NVIDIA 4060 GPU, batch size 32) to ensure fairness. Inference time measurements include forward propagation but exclude data loading and are averaged over 1000 samples. Qwen model inference times are based on measured 8-bit quantized implementations.

Table 10. Efficiency-Performance Comparison on the CMU-MOSEI Test Set. $\uparrow$ indicates that higher values are better, whereas $\downarrow$ denotes that lower values are better. The best result for each metric is shown in bold.

Model	Params (M)	Infer. Time (ms)	MAE $\downarrow$	Corr $\uparrow$
TFN	8.2	6.1	0.593	0.700
LMF	5.4	4.8	0.623	0.677
MuT	18.7	11.2	0.580	0.703
MISA	15.3	9.4	0.568	0.724
SKEAFN	14.1	10.7	0.540	0.763
Qwen-Text	7000.0	142.0	0.528	0.771
Qwen-Fusion	7020.0	156.0	0.519	0.775
MSF-TGRA	12.6	8.3	0.512	0.763

4.8. Case Study: Complementary Role of Multimodal Information

To qualitatively demonstrate the effectiveness of the text-guided multimodal mechanism, we selected samples from the CMU-MOSEI test set with textual sentiment entropy greater than 0.8 (ambiguous) or whose text polarity contradicts the labels (sarcastic). Two representative cases were randomly chosen for detailed analysis.

Case 1:

Ambiguous Textual Sentiment (Sample ID: MOSEI_test_1247).

- Ground Truth: Negative (−2)
- Text Content: “I mean, it’s fine, I guess.”
- Text-Only Prediction: Neutral (0.1)
- MSF-TGRA Prediction: Negative (−1.8) ✓

Analysis 1:

The text lacks explicit emotional words and has a neutral semantic tendency, causing the text-only model to misjudge the sentiment polarity. However, the visual modality shows slight frowning (Action Unit AU4, brow lowering, activation = 0.7) and reduced facial engagement (AU12, lip corner pulling, activation = 0.2), while the audio modality exhibits low energy (mean F0 = 112 Hz, below the speaker’s average of 145 Hz) and a flat pitch contour (range = 23 Hz). Under the “weak-emotion” guidance from textual semantics, the TGMHA module assigns higher weights to emotion-relevant visual segments (micro-expression frames) and acoustic segments (low-energy regions), achieving fine-grained cross-modal alignment. Meanwhile, the OFDRA module suppresses redundant temporal information (e.g., static frames and steady speech segments), retaining only key behavioral cues. The multi-scale fusion mechanism plays a critical role: the 3×3 kernel captures instantaneous frowning actions, the 5×5 kernel models gradual facial transitions from neutral to negative, and the 7×7 kernel confirms sustained low engagement, with all three scales consistently indicating negative sentiment. This case demonstrates that, when textual sentiment is ambiguous, the text-guided multimodal cues can effectively complement the text and correct mispredictions.

Case 2:

Conflict Between Textual and Non-Textual Modalities (Sarcasm) (Sample ID: MOSEI_test_3891).

- Ground Truth: Negative (−2)
- Text Content: “Yeah, that was really great.”
- Text-Only Prediction: Positive (+2.1)
- MSF-TGRA Prediction: Negative (−1.6) ✓

Analysis 2:

The text explicitly conveys a positive sentiment, causing the text-only model to misclassify the sample as positive. However, the ground truth is negative, reflecting sarcastic expression. MSF-TGRA successfully identifies this discrepancy through non-textual cues: the audio modality exhibits decreasing pitch (F0 drops from 180 Hz to 140 Hz), slowed speech rate (0.8× normal speed), and elongated voice onset time (VOT extended by 40%), while the visual modality shows disengaged facial expressions (50% reduction in eye contact, slight downward lip movement (AU15, lip corner depressing, activation = 0.6)). The multi-scale design of TGMHA captures both local temporal inconsistencies (negative audiovisual cues corresponding to the textual positive word “great”) and global cross-modal dependencies (overall negative tone in prosody and facial expression), detecting sentiment incongruence between text and other modalities. Consequently, the model avoids over-reliance on surface textual polarity, producing the correct negative prediction.

Discussions:

These cases demonstrate that, although textual information dominates the overall performance on CMU-MOSEI, non-textual modalities can provide complementary and discriminative emotional cues in specific but practically relevant scenarios, such as ambiguous or sarcastic expressions. The proposed text-guided mechanism is not intended to replace textual semantics; rather, it selectively activates informative visual and acoustic cues when textual evidence is insufficient or ambiguous. This qualitative analysis supports the quantitative results and validates the rationale for employing a task-specific multimodal architecture in text-guided sentiment analysis scenarios.

4.9. Modality-Dominant Analysis

We force the fusion pipeline to rely on a single modality by masking the other two (after projection), aiming to quantify the relative importance of each modality. Table 11 shows that text-guided fusion achieves the lowest MAE and highest correlation, while visual or acoustic dominance degrades performance sharply. This outcome validates the design choice of using text as the guiding modality: textual tokens carry the most direct sentiment-related semantics, whereas visual and acoustic signals are more susceptible to speaker- and context-specific variations. Consequently, text-guided attention improves accuracy by filtering out irrelevant variability in the other streams.

Table 11. Modality-dominant experiment on CMU-MOSEI. ↑ indicates that higher values are better, whereas ↓ denotes that lower values are better.

Variant	MAE ↓	Corr ↑
Text-dominant	0.624	0.642
Visual-dominant	0.815	0.167
Acoustic-dominant	0.814	0.224

4.10. Visualization of Text-Guided Attention

Figure 4 illustrates how textual semantics guide the attention mechanism across modalities. In this example, the sentiment-bearing token “excellent” activates high attention weights over both the corresponding smiling video frame (visual modality) and the segment with elevated pitch energy (acoustic modality). This demonstrates that text features effectively highlight semantically relevant visual and acoustic cues.

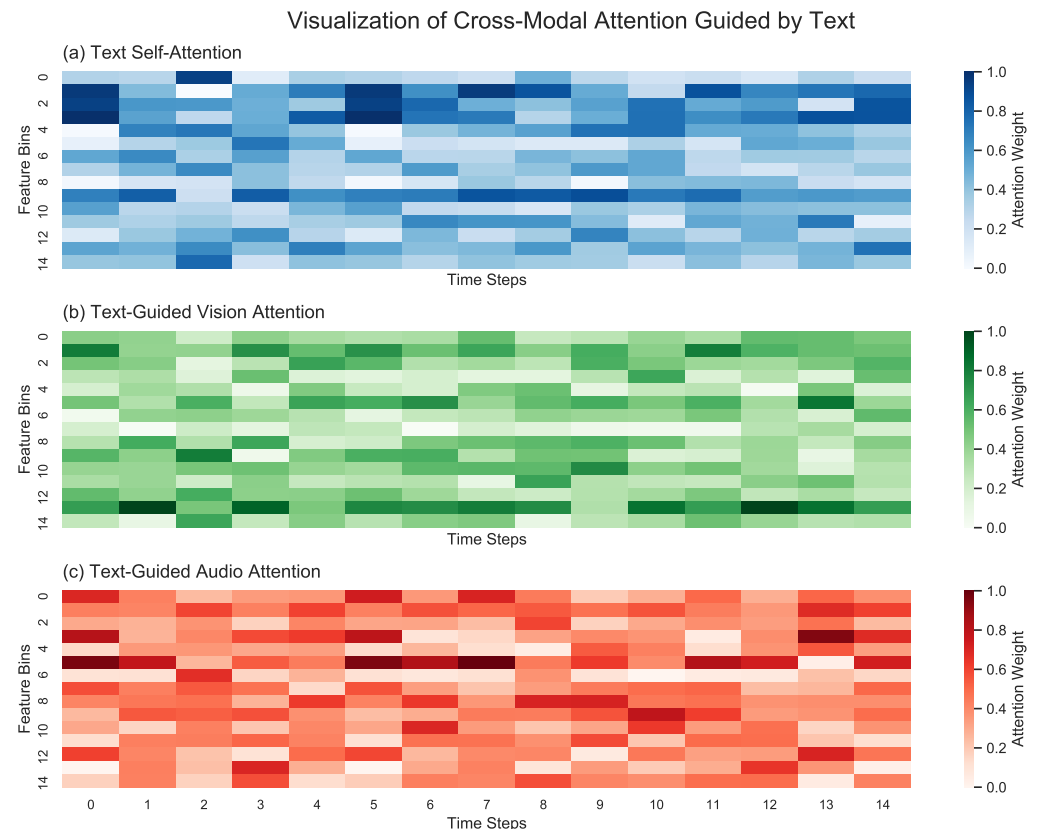


Figure 4. Text-guided cross-modal attention heatmaps. **(a)** Text self-attention highlights sentiment-bearing tokens and internal dependencies. **(b)** Text-guided visual attention focuses on semantically relevant video frames. **(c)** Text-guided acoustic attention emphasizes aligned audio segments. Brighter regions indicate higher attention scores.

The visualization includes three heatmaps: (a) text self-attention, (b) text-guided visual attention, and (c) text-guided acoustic attention. Brighter values indicate higher attention weights. The text self-attention map (a) shows internal dependencies among textual features, where the vertical axis represents text feature dimensions and the horizontal axis denotes text time steps. The text-guided visual attention map (b) depicts which video-frame features are emphasized at each text time step, with the vertical axis representing visual feature dimensions and the horizontal axis representing frame indices. Similarly, the text-guided acoustic map (c) highlights audio-segment features influenced by text semantics. These qualitative results support the claim in Section 3 that textual information serves as a stable anchor for selecting sentiment-relevant cues from non-verbal modalities, thereby enhancing cross-modal alignment.

5. Conclusions and Future Work

5.1. Conclusions

In this article, we proposed MSF-TGRA, a text-guided multi-scale fusion model for multimodal sentiment analysis (MSA). The model addresses two critical challenges in MSA: excessive spatiotemporal redundancy within individual modalities and insufficient fine-grained cross-modal alignment. To overcome these limitations, we introduced two core components: the Overlapping Feature Dimensionality Reduction Attention (OFDRA) module, which effectively mitigates intra-modality feature redundancy, and the Text-Guided Multi-Scale Hybrid Attention (TGMHA) module, which establishes discriminative cross-modal alignment under semantic guidance.

Comprehensive evaluations on four benchmark datasets—CMU-MOSI, CMU-MOSEI, CH-SIMS, and MELD—demonstrate that MSF-TGRA consistently outperforms state-of-the-art baselines in terms of MAE and correlation metrics. Comparative analyses with LLMs (e.g., Qwen variants) reveal that our lightweight architecture achieves competitive performance while requiring $556\times$ fewer parameters and $17\times$ lower inference latency.

This work does not aim to compete with LLMs on text-only sentiment analysis tasks. Instead, it focuses on scenarios where multimodal data are available and non-textual modalities provide complementary affective cues, such as visual micro-expressions and acoustic prosodic variations. Through systematic case analyses, we show that text-guided multimodal mechanisms can improve predictive accuracy in challenging settings, including ambiguous emotional expressions and the presence of sarcasm. The primary contribution of this study lies not in achieving substantial absolute performance gains, but in clarifying the complementary role of multimodal information under text-guided conditions, as well as the associated design principles that govern its effective integration.

In summary, this work makes three key contributions. First, we introduce a text-guided multi-scale fusion model that achieves an effective balance between computational efficiency and predictive performance. Second, we provide empirical evidence that textual semantics can act as a reliable anchor for aligning information across modalities in sentiment analysis tasks. Third, through systematic experimentation, we demonstrate that well-designed lightweight architectures can be competitive with—and in some cases outperform—larger LLM-based approaches for specific MSA applications.

Based on our comprehensive experiments, we identify two specific scenarios where multimodal fusion demonstrates clear advantages over text-only models: (1) ambiguous textual sentiment (Case Study, Section 4.8), where non-textual behavioral cues disambiguate neutral or vague linguistic expressions; and (2) sarcastic or incongruent expressions, where audiovisual signals contradict surface-level textual polarity and enable correct sentiment inference. These findings provide actionable guidance for practitioners: multimodal architectures are most justified when deployment conditions involve unreliable or ambiguous textual input, and text-only models may suffice when high-quality transcripts are consistently available.

5.2. Limitations

Despite the promising results of the proposed method, several limitations should be acknowledged as current constraints of the present framework.

First, the model's generalization to specialized domains remains uncertain. While it performs well on product review and social video datasets such as CMU-MOSI and CMU-MOSEI, its applicability to domains such as clinical or financial sentiment analysis still requires further validation. In addition, the performance gains are context-dependent: when textual information is clear and sufficient, the marginal benefit of the multimodal design may be limited, whereas multimodal fusion becomes more valuable under ambiguous sentiment conditions.

Second, the current architecture relies on the assumption that text serves as a reliable semantic anchor. This assumption is reasonable when textual input is of sufficient quality, such as human-generated transcripts or high-accuracy automatic speech recognition (ASR) output. However, in real-world deployment, text quality may degrade due to ASR errors, code-mixed utterances, highly informal language, or domain-specific jargon. In such cases, the text-guided attention mechanism may propagate unreliable semantic signals to the visual and acoustic branches. A related issue arises with sarcasm, irony, and other forms of figurative language, where surface-level textual polarity is misleading: although the case study in Section 4.8 suggests that non-textual cues can help in some instances, the present

architecture does not perform dedicated pragmatic reasoning or sarcasm detection, and performance may degrade when both textual and non-textual cues are ambiguous. Future work should explore adaptive anchor selection, confidence-aware guidance mechanisms, or dedicated sarcasm detection modules to address these scenarios.

Third, the current implementation relies on fixed pre-extracted features (GloVe, Facet, and COVAREP), which limits the ability of the model to benefit from recent advances in self-supervised representation learning. These extractors were retained to ensure strict comparability with all baseline methods under identical input conditions. Although the proposed architecture is compatible with more modern encoders such as RoBERTa [15], DeBERTa [67], wav2vec 2.0 [68], or HuBERT [65], adopting them in the current study would make it harder to separate gains from improved input representations from gains due to the fusion architecture itself.

5.3. Future Work

Future research could proceed along the following directions.

First, model efficiency and adaptability could be improved by exploring lightweight attention mechanisms and parameter-efficient training methods, such as low-rank adaptation (LoRA) or quantization. Additionally, integrating structured reasoning modules to incorporate external commonsense knowledge may further enhance the model's contextual understanding.

Second, extending the model to more diverse real-world scenarios—including multilingual and cross-domain sentiment analysis, context-dependent modality selection, and downstream tasks such as sarcasm detection or empathetic dialogue—would help evaluate its generalization capability across varied emotional environments.

Finally, investigating hybrid architectures that combine LLMs with the proposed fusion framework is a promising direction. Leveraging the powerful semantic understanding of LLMs as a text backbone, while retaining our lightweight multimodal fusion modules to process non-textual information, could achieve a better trade-off between performance and computational efficiency.

Author Contributions: Conceptualization, S.K. and J.Z.; methodology, S.K.; software, S.K.; validation, S.K. and W.L.; writing—original draft preparation, S.K.; writing—review and editing, J.Z. and P.A.; supervision, J.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The raw data supporting the conclusions of this article will be made available by the authors on request.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Zhang, C.; Yang, Z.; He, X.; Deng, L. Multimodal intelligence: Representation learning, information fusion, and applications. *IEEE J. Sel. Top. Signal Process.* **2020**, *14*, 478–493. [\[CrossRef\]](#)
2. Baltrušaitis, T.; Ahuja, C.; Morency, L.P. Multimodal machine learning: A survey and taxonomy. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**, *41*, 423–443. [\[CrossRef\]](#)
3. Akhtar, M.S.; Chauhan, D.S.; Ghosal, D.; Poria, S.; Ekbal, A.; Bhattacharyya, P. Multi-task learning for multi-modal emotion recognition and sentiment analysis. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2019), Minneapolis, MN, USA, 2–7 June 2019; pp. 370–379.
4. Esuli, A.; Sebastiani, F. SentiWordNet: A publicly available lexical resource for opinion mining. In Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'06), Genoa, Italy, 22–28 May 2006; pp. 417–422.
5. Miller, G.A. WordNet: A lexical database for English. *Commun. ACM* **1995**, *38*, 39–41. [\[CrossRef\]](#)

6. Baccianella, S.; Esuli, A.; Sebastiani, F. SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC'10), Valletta, Malta, 17–23 May 2010; pp. 2200–2204.
7. Cambria, E.; Speer, R.; Havasi, C.; Hussain, A. SenticNet: A publicly available semantic resource for opinion mining. In Proceedings of the AAAI Fall Symposium: Commonsense Knowledge, Arlington, VA, USA, 11–13 November 2010; pp. 14–18.
8. Cambria, E.; Hussain, A. Sentic Computing. *Cogn. Comput.* **2015**, *7*, 183–185. [[CrossRef](#)]
9. Zhang, S.; Chen, C.; Chen, J.; Li, Y.F.; Wu, Y.; Li, M.; Zhu, C. Combining cross-modal knowledge transfer and semi-supervised learning for speech emotion recognition. *Knowl.-Based Syst.* **2021**, *229*, 107340. [[CrossRef](#)]
10. Wu, T.; Peng, J.; Zhang, W.; Zhang, H.; Tan, S.; Yi, F.; Ma, C.; Huang, Y. Video sentiment analysis with bimodal information-augmented multi-head attention. *Knowl.-Based Syst.* **2022**, *235*, 107676. [[CrossRef](#)]
11. Wang, R.; Hao, Y.; Yu, Q.; Chen, M.; Humar, I.; Fortino, G. Depression analysis and recognition based on functional near-infrared spectroscopy. *IEEE J. Biomed. Health Inf.* **2021**, *25*, 4289–4299. [[CrossRef](#)]
12. Rahman, W.; Hasan, M.K.; Lee, S.; Zadeh, A.; Mao, C.; Morency, L.P.; Hoque, E. Integrating multimodal information in large pretrained transformers. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL '20), Online, 5–10 July 2020; pp. 2359–2369.
13. Tsai, Y.H.H.; Liang, P.P.; Zadeh, A.; Morency, L.P.; Salakhutdinov, R. Learning factorized multimodal representations. In Proceedings of the 7th International Conference on Learning Representations (ICLR 2019), New Orleans, LA, USA, 6–9 May 2019. Available online: <https://openreview.net/forum?id=rygqqsA9KX> (accessed on 20 May 2026).
14. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. 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 (NAACL-HLT 2019), Minneapolis, MN, USA, 2–7 June 2019; pp. 4171–4186.
15. Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; Stoyanov, V. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv* **2019**, arXiv:1907.11692.
16. Han, W.; Chen, H.; Poria, S. Improving multimodal fusion with hierarchical mutual information maximization for multimodal sentiment analysis. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP 2021), Punta Cana, Dominican Republic, 7–11 November 2021; pp. 9180–9192.
17. Ridnik, T.; Ben-Baruch, E.; Zamir, N.; Noy, A.; Friedman, I.; Protter, M.; Zelnik-Manor, L. Asymmetric loss for multi-label classification. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montréal, QC, Canada, 10–17 October 2021; pp. 82–91.
18. Hazarika, D.; Zimmermann, R.; Poria, S. MISA: Modality-invariant and modality-specific representations for multimodal sentiment analysis. In Proceedings of the 28th ACM International Conference on Multimedia (MM '20), Seattle, WA, USA, 12–16 October 2020; pp. 1122–1131.
19. Cai, Y.; Cai, H.; Wan, X. Multi-modal sarcasm detection in Twitter with hierarchical fusion model. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL '19), Florence, Italy, 28 July–2 August 2019; pp. 2506–2515.
20. Xu, N.; Mao, W. MultiSentiNet: A deep semantic network for multimodal sentiment analysis. In Proceedings of the 2017 ACM on Conference on Information and Knowledge Management (CIKM '17), Singapore, 6–10 November 2017; pp. 2399–2402.
21. Pan, H.; Lin, Z.; Fu, P.; Qi, Y.; Wang, W. Modeling intra and inter-modality incongruity for multi-modal sarcasm detection. In Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2020, Online, 16–20 November 2020; pp. 1383–1392.
22. Xu, N.; Zeng, Z.; Mao, W. Reasoning with multimodal sarcastic tweets via modeling cross-modality contrast and semantic association. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL '20), Online, 5–10 July 2020; pp. 3777–3786.
23. Liu, Z.; Shen, Y.; Lakshminarasimhan, V.B.; Liang, P.P.; Zadeh, A.; Morency, L.P. Efficient low-rank multimodal fusion with modality-specific factors. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL '18), Melbourne, Australia, 15–20 July 2018; pp. 2247–2256.
24. Hou, M.; Tang, J.; Zhang, J.; Kong, W.; Zhao, Q. Deep multimodal multilinear fusion with high-order polynomial pooling. In Proceedings of the 33rd International Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, BC, Canada, 8–14 December 2019; pp. 12156–12166.
25. Jin, T.; Huang, S.; Li, Y.; Zhang, Z. Dual low-rank multimodal fusion. In Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2020, Online, 16–20 November 2020; pp. 377–387.
26. Mai, S.; Hu, H.; Xing, S. Modality to modality translation: An adversarial representation learning and graph fusion network for multimodal fusion. In Proceedings of the 34th AAAI Conference on Artificial Intelligence (AAAI-20), New York, NY, USA, 7–12 February 2020; pp. 164–172.
27. Zhu, C.; Chen, M.; Zhang, S.; Sun, C.; Liang, H.; Liu, Y.; Chen, J. SKEAFN: Sentiment Knowledge Enhanced Attention Fusion Network for multimodal sentiment analysis. *Inf. Fusion* **2023**, *100*, 101958. [[CrossRef](#)]

28. Yang, K.; Xu, H.; Gao, K. CM-BERT: Cross-modal BERT for text-audio sentiment analysis. In Proceedings of the 28th ACM International Conference on Multimedia (MM '20), Seattle, WA, USA, 12–16 October 2020; pp. 521–528.
29. Zhao, X.; Chen, Y.; Li, W.; Gao, L.; Tang, B. MAG+: An extended multimodal adaptation gate for multimodal sentiment analysis. In Proceedings of the 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, 22–27 May 2022; pp. 4753–4757.
30. Chen, Q.; Huang, G.; Wang, Y. The weighted cross-modal attention mechanism with sentiment prediction auxiliary task for multimodal sentiment analysis. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2022**, *30*, 2689–2695. [CrossRef]
31. Salur, M.U.; Aydın, İ. A soft voting ensemble learning-based approach for multimodal sentiment analysis. *Neural Comput. Appl.* **2022**, *34*, 18391–18406. [CrossRef]
32. Wang, Y.; Shen, Y.; Liu, Z.; Liang, P.P.; Zadeh, A.; Morency, L.P. Words can shift: Dynamically adjusting word representations using nonverbal behaviors. In Proceedings of the 33rd AAAI Conference on Artificial Intelligence (AAAI'19), Honolulu, HI, USA, 27 January–1 February 2019; pp. 7216–7223.
33. Chen, R.; Zhou, W.; Li, Y.; Zhou, H. Video-based cross-modal auxiliary network for multimodal sentiment analysis. *IEEE Trans. Circuits Syst. Video Technol.* **2022**, *32*, 8703–8716. [CrossRef]
34. Mai, S.; Xing, S.; He, J.; Zeng, Y.; Hu, H. Multimodal graph for unaligned multimodal sequence analysis via graph convolution and graph pooling. *ACM Trans. Multimed. Comput. Commun. Appl.* **2023**, *19*, 54. [CrossRef]
35. Zadeh, A.; Liang, P.P.; Poria, S.; Cambria, E.; Morency, L.P. 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 (ACL '18), Melbourne, Australia, 15–20 July 2018; pp. 2236–2246.
36. Ghorbanali, A.; Sohrabi, M.K.; Yaghmaee, F. Ensemble transfer learning-based multimodal sentiment analysis using weighted convolutional neural networks. *Inf. Process. Manag.* **2022**, *59*, 102929. [CrossRef]
37. Zhu, T.; Li, L.; Yang, J.; Zhao, S.; Liu, H.; Qian, J. Multimodal sentiment analysis with image-text interaction network. *IEEE Trans. Multimed.* **2023**, *25*, 3375–3385. [CrossRef]
38. Nguyen, H.L.; Vu, D.T.; Jung, J.J. Knowledge graph fusion for smart systems: A survey. *Inf. Fusion* **2020**, *61*, 56–70. [CrossRef]
39. Wu, Z.; Pan, S.; Chen, F.; Long, G.; Zhang, C.; Yu, P.S. A comprehensive survey on graph neural networks. *IEEE Trans. Neural Netw. Learn. Syst.* **2021**, *32*, 4–24. [CrossRef] [PubMed]
40. Han, W.; Chen, H.; Gelbukh, A.; Zadeh, A.; Morency, L.P.; Poria, S. Bi-bimodal modality fusion for correlation-controlled multimodal sentiment analysis. In Proceedings of the 2021 International Conference on Multimodal Interaction (ICMI '21), Montréal, QC, Canada, 18–22 October 2021; pp. 6–15.
41. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef]
42. Zadeh, A.; Chen, M.; Poria, S.; Cambria, E.; Morency, L.P. Tensor Fusion Network for multimodal sentiment analysis. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP 2017), Copenhagen, Denmark, 7–11 September 2017; pp. 1103–1114.
43. Sun, Z.; Sarma, P.K.; Sethares, W.A.; Liang, Y. Learning relationships between text, audio, and video via deep canonical correlation for multimodal language analysis. In Proceedings of the 34th AAAI Conference on Artificial Intelligence (AAAI-20), New York, NY, USA, 7–12 February 2020; pp. 8992–8999.
44. Zadeh, A.; Liang, P.P.; Poria, S.; Vij, P.; Cambria, E.; Morency, L.P. Multi-attention recurrent network for human communication comprehension. In Proceedings of the 32nd AAAI Conference on Artificial Intelligence (AAAI'18), New Orleans, LA, USA, 2–7 February 2018; pp. 5642–5649.
45. Jin, W.; Zhao, B.; Zhang, L.; Liu, C.; Yu, H. Back to common sense: Oxford dictionary descriptive knowledge augmentation for aspect-based sentiment analysis. *Inf. Process. Manag.* **2023**, *60*, 103260. [CrossRef]
46. Chauhan, D.S.; Dhanush, S.R.; Ekbal, A.; Bhattacharyya, P. Sentiment and emotion help sarcasm? A multi-task learning framework for multi-modal sarcasm, sentiment and emotion analysis. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL '20), Online, 5–10 July 2020; pp. 4351–4360.
47. Gupta, S.; Shah, A.; Shah, M.; Syiemlieh, L.; Maurya, C. FiLMing multimodal sarcasm detection with attention. In Proceedings of the 28th International Conference on Neural Information Processing (ICONIP 2021), Sanur, Bali, Indonesia, 8–12 December 2021; pp. 178–186.
48. OpenAI. GPT-4 Technical Report. *arXiv* **2023**, arXiv:2303.08774.
49. Alayrac, J.B.; Donahue, J.; Luc, P.; Miech, A.; Barr, I.; Hasson, Y.; Lenc, K.; Mensch, A.; Millicah, K.; Reynolds, M.; et al. Flamingo: A visual language model for few-shot learning. In Proceedings of the 36th International Conference on Neural Information Processing Systems (NIPS'22), New Orleans, LA, USA, 28 November–9 December 2022; pp. 23716–23736.
50. Wei, J.; Bosma, M.; Zhao, V.Y.; Guu, K.; Yu, A.W.; Lester, B.; Du, N.; Dai, A.M.; Le, Q.V. Finetuned Language Models are Zero-Shot Learners. In Proceedings of the 10th International Conference on Learning Representations (ICLR 2022), Online, 25–29 April 2022. Available online: <https://openreview.net/forum?id=gEZrGCozdqR> (accessed on 20 May 2026).

51. Sanh, V.; Webson, A.; Raffel, C.; Bach, S.H.; Sutawika, L.; Alyafeai, Z.; Chaffin, A.; Stiegler, A.; Le Scao, T.; Raja, A.; et al. Multitask Prompted Training Enables Zero-Shot Task Generalization. In Proceedings of the 10th International Conference on Learning Representations (ICLR 2022), Online, 25–29 April 2022. Available online: <https://openreview.net/forum?id=9Vrb9D0WI4> (accessed on 20 May 2026).
52. Zadeh, A.; Liang, P.P.; Mazumder, N.; Poria, S.; Cambria, E.; Morency, L.P. Memory fusion network for multi-view sequential learning. In Proceedings of the 32nd AAAI Conference on Artificial Intelligence and 30th Innovative Applications of Artificial Intelligence Conference and 8th AAAI Symposium on Educational Advances in Artificial Intelligence (AAAI'18/IAAI'18/EAAI'18), New Orleans, LA, USA, 2–7 February 2018; pp. 5634–5641.
53. Hu, E.J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W. LoRA: Low-Rank Adaptation of Large Language Models. *arXiv* **2021**, arXiv:2106.09685.
54. Sanh, V.; Debut, L.; Chaumond, J.; Wolf, T. DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter. *arXiv* **2019**, arXiv:1910.01108.
55. Zhang, H.; Zhang, Y.; Xu, G.; Zheng, J.; Zhao, M. Multi-perspective text-guided multimodal fusion network for brain tumor segmentation. In Proceedings of the 7th Chinese Conference on Pattern Recognition and Computer Vision (PRCV 2024), Urumqi, China, 18–20 October 2024; pp. 299–313.
56. Wu, X.; Zhao, K.; Huang, Q.; Wang, Q.; Yang, Z.; Hao, G. MISL: Multi-grained image-text semantic learning for text-guided image inpainting. *Pattern Recognit.* **2024**, *145*, 109961. [[CrossRef](#)]
57. Liao, X.; Ke, X.; Liu, Q.; Li, G. Text-guided enhanced transformer fusion for multimodal sentiment analysis. *J. Shanghai Jiaotong Univ. (Sci.)* **2025**. [[CrossRef](#)]
58. Liu, M.; Dan, J.; Lu, Z.; Yu, Y.; Li, Y.; Li, X. CM-UNet: Hybrid CNN-Mamba UNet for remote sensing image semantic segmentation. *arXiv* **2024**, arXiv:2405.10530.
59. Lou, M.; Zhang, S.; Zhou, H.Y.; Yang, S.; Wu, C.; Yu, Y. TransXNet: Learning both global and local dynamics with a dual dynamic token mixer for visual recognition. *IEEE Trans. Neural Netw. Learn. Syst.* **2025**, *36*, 11534–11547. [[CrossRef](#)]
60. Pennington, J.; Socher, R.; Manning, C.D. GloVe: Global vectors for word representation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP 2014), Doha, Qatar, 25–29 October 2014; pp. 1532–1543.
61. Degottex, G.; Kane, J.; Drugman, T.; Raitio, T.; Scherer, S. COVAREP—A collaborative voice analysis repository for speech technologies. In Proceedings of the 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Florence, Italy, 4–9 May 2014; pp. 960–964.
62. Baltrušaitis, T.; Zadeh, A.; Lim, Y.C.; Morency, L.P. OpenFace 2.0: Facial behavior analysis toolkit. In Proceedings of the 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018), Xi'an, China, 15–19 May 2018; pp. 59–66.
63. Yu, W.; Xu, H.; Yuan, Z.; Wu, J. Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis. In Proceedings of the 35th AAAI Conference on Artificial Intelligence (AAAI-21), Online, 2–9 February 2021; pp. 10790–10797.
64. Poria, S.; Hazarika, D.; Majumder, N.; Naik, G.; Cambria, E.; Mihalcea, R. MELD: A multimodal multi-party dataset for emotion recognition in conversations. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL '19), Florence, Italy, 28 July–2 August 2019; pp. 527–536.
65. Hsu, W.N.; Bolte, B.; Tsai, Y.H.H.; Lakhota, K.; Salakhutdinov, R.; Mohamed, A. HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2021**, *29*, 3451–3460. [[CrossRef](#)]
66. Radford, A.; Kim, J.W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. Learning Transferable Visual Models from Natural Language Supervision. In Proceedings of the 38th International Conference on Machine Learning (ICML 2021), Online, 18–24 July 2021; pp. 8748–8763.
67. He, P.; Liu, X.; Gao, J.; Chen, W. DeBERTa: Decoding-Enhanced BERT with Disentangled Attention. In Proceedings of the International Conference on Learning Representations (ICLR 2021), Vienna, Austria, 3–7 May 2021; pp. 3986–4006.
68. Baevski, A.; Zhou, Y.; Mohamed, A.; Auli, M. wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. In Proceedings of the 34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, BC, Canada, 6–12 December 2020; pp. 12449–12460.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.