Computers, Materials & Continua

Tech Science Press

Doi:10.32604/cmc.2025.068126

ARTICLE

Enhanced Multimodal Sentiment Analysis via Integrated Spatial Position Encoding and Fusion Embedding

Chenquan Gan[1,2,*], Xu Liu[1], Yu Tang[2], Xianrong Yu[3], Qingyi Zhu[1] and Deepak Kumar Jain[4]

[1]School of Cyber Security and Information Law, Chongqing University of Posts and Telecommunications, Chongqing, 400065, China
[2]School of Communication and Information Engineering, Chongqing University of Posts and Telecommunications, Chongqing, 400065, China
[3]Jiangxi Provincial Key Laboratory of Electronic Data Control and Forensics (Jiangxi Police College), Nanchang, 330100, China
[4]Key Laboratory of Intelligent Control and Optimization for Industrial Equipment of Ministry of Education, Dalian University of Technology, Dalian, 116024, China

*Corresponding Author: Chenquan Gan. Email: gancq@cqupt.edu.cn

Received: 21 May 2025; Accepted: 26 August 2025; Published: 23 October 2025

ABSTRACT: Multimodal sentiment analysis aims to understand emotions from text, speech, and video data. However, current methods often overlook the dominant role of text and suffer from feature loss during integration. Given the varying importance of each modality across different contexts, a central and pressing challenge in multimodal sentiment analysis lies in maximizing the use of rich intra-modal features while minimizing information loss during the fusion process. In response to these critical limitations, we propose a novel framework that integrates spatial position encoding and fusion embedding modules to address these issues. In our model, text is treated as the core modality, while speech and video features are selectively incorporated through a unique position-aware fusion process. The spatial position encoding strategy preserves the internal structural information of speech and visual modalities, enabling the model to capture localized intra-modal dependencies that are often overlooked. This design enhances the richness and discriminative power of the fused representation, enabling more accurate and context-aware sentiment prediction. Finally, we conduct comprehensive evaluations on two widely recognized standard datasets in the field—CMU-MOSI and CMU-MOSEI to validate the performance of the proposed model. The experimental results demonstrate that our model exhibits good performance and effectiveness for sentiment analysis tasks.

KEYWORDS: Multimodal sentiment analysis; spatial position encoding; fusion embedding; feature loss reduction

1 Introduction

With the increasing prevalence of social media, individuals are progressively inclined to utilize diverse forms of data to articulate their ideological perspectives and emotional sentiments on these platforms. Early sentiment analysis was limited to mining and analyzing emotional tendencies and stances related to specific subjects using single-modal data, such as text [1]. This approach overlooks the rich emotional information contained in various modalities, such as voice and video, complicating the accurate analysis of emotional tendencies in many contexts. Consequently, the scope of multimodal sentiment analysis encompasses a wide range of information domains, including text, audio, and video [2]. The field has gained attention due to its capacity to process complex data [3]. Moreover, its ability to extract authentic emotions and opinions from multimodal data facilitates applications in practical areas, such as social recommendation [4], trust management [5], and mental health [6].

Multimodal data encompasses an extensive and complex range of information dimensions. Taking text as an example, simple phrases such as *"I didn't expect you at this time"* may imply anger in an isolated text, while *"You're a genius"* directly conveys surprise, and *"Why did you come to me again?"* may imply dissatisfaction or disgust. However, when we introduce non-textual cues such as speech and video as auxiliary analysis tools, the boundaries of these emotional interpretations become blurred and enriched. For example, the first phrase, with a cheerful tone and a video background of hugs, may transform into joy at the visit of an intimate friend. If a self-deprecating or regretful expression accompanies the second phrase, it may express self-blame for the wrong behavior rather than pure praise. In the scene where the other person appears with a smiling face, the third phrase is likely to be transformed into an unexpected joy at the visit of the person you admire. As shown in Fig. 1, this example of multimodal sentiment analysis vividly demonstrates the power of the comprehensive study. In [7], the combination of the three modal data achieves 80.58% and 79.63% accuracy, respectively, which far exceeds the accuracy of unimodal. As a result, integrating textual, speech, and video modal information can enhance the performance of sentiment analysis tasks.

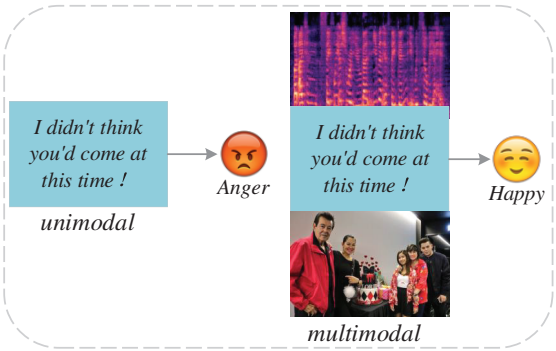

Figure 1: An example of unimodal and multimodal results

The multimodal fusion strategy can be refined into two mainstream branches: model-independent [8] and model-based [9]. Under the model-independent fusion framework, the strategy is further divided into early fusion and late fusion. Early fusion focuses on the feature level, capturing the interaction information in the initial stage by integrating the vectors of various modalities. And late-stage fusion focuses on the decision-making level, comprehensively considering the preliminary judgment results of each modality. Among them, LMF (Latent Multimodal Fusion) [10] is a typical example, which excels in mining hidden connections between low-level features. Although such methods can address most fusion challenges, excessive reliance on low-level features may limit their ability to explore complex relationships between data in depth.

In contrast, model-based methods, especially the multimodal fusion strategy combining machine learning and deep learning, aim to reveal deeper levels of interdependence and synergy between modalities. For example, Long Short Term Memory (LSTM) [11] demonstrates powerful temporal modeling capabilities in multimodal sentiment analysis by introducing a time dimension. The introduction of the Transformer framework [12] provides an efficient and flexible architecture for integrating multiple single-modal information. Recently, attention mechanisms have emerged as a key solution to long-term dependency issues in sequence modeling [13]. It enhances the model's ability to capture important information by dynamically adjusting weight allocation. However, the limitation of methods such as cross transformers is that they often assume equal contributions from all modalities and fail to fully consider the unique value of key modalities such as text, resulting in performance bottlenecks [14]. To overcome this limitation, gate control mechanisms have emerged, which dynamically adjust the weight ratios of different modalities in the fusion process,

achieving more refined control. But this advantage is also accompanied by a increase in computing resource demand [15].

Given the varying importance of different modalities, how to skillfully integrate multimodal data and maximize the utilization of extracted rich features has become the core challenge in improving the accuracy of sentiment analysis. Unfortunately, there are still shortcomings in the current fusion strategy, especially the phenomenon of feature loss. To overcome this bottleneck, we design a new multimodal sentiment analysis framework. This framework seamlessly integrates information from three modalities: text, speech, and video. Through a carefully optimized fusion mechanism, it aims to utilize the extracted features more comprehensively and can improve sentiment analysis accuracy in our experiments. Our model captures the core features of each modality, reducing information loss during the fusion process and helping to better preserve key information. It is particularly worth mentioning that, in response to the common problem of feature loss in speech and video modalities, we introduce a spatial position encoding strategy. However, most existing approaches treat speech and video features as simple sequences, neglecting their inherent internal structure. For example, speech features commonly adopt time–frequency representations (e.g., spectrograms), which naturally form a two-dimensional grid. Flattening these into one-dimensional sequences discards frequency-locality information critical for sentiment cues such as pitch variation or energy concentration. Similarly, video frames possess spatial layouts that are essential for interpreting expressions. To address this, we propose spatial position encoding to preserve these structural properties, thereby mitigating feature loss during fusion.

The main contributions can be summarized as follows:

1) We develop a multimodal sentiment analysis model that combines fusion embedding and spatial position encoding. This model uses text as the core modality and integrates it with speech and video, ensuring the ultimate preservation of key features in each modality.
2) By implementing spatial position encoding strategies for speech and video, we reduce the loss of spatial information in the network processing of these two modalities, thereby promoting the efficient utilization of internal features of the modalities.
3) To verify the performance of the model, we conduct comprehensive tests on the standard datasets CMU-MOSI and CMU-MOSEI, and the experimental results demonstrate the performance and effectiveness of the model in the field of sentiment analysis.

2 Related Work

Currently, the mainstream methods in the field of unimodal sentiment analysis cover text sentiment parsing based on lexicon and machine learning techniques, image sentiment recognition using convolutional neural network (CNN) and VGG model [16], and speech sentiment analysis based on support vector machine (SVM) [17]. However, with the ever-changing advancement of Internet technology and the increasingly rich and diverse data forms, single-modal sentiment analysis methods are not capable of capturing and integrating the complex inter-modal correlation information, which tends to limit the comprehensiveness and accuracy of the analysis results [18]. Further, in the face of complex and changing contextual environments, unimodal analyses also reveal computational inefficiencies and poor adaptability.

In view of this, multimodal sentiment analysis, which integrates multiple data modalities, has emerged as a new trend in the field of sentiment intelligence. Zadeh et al. [19] pioneered the introduction of tensor fusion networks (TFNs), which enable seamless integration of information from unimodal to bimodal and even trimodal. However, with the cumulative expansion of the matrix product during computation, the dimensionality of the feature vectors climbs dramatically, placing a heavy burden on model training. Deeply inspired by TFNs, the work innovatively introduced a low-rank weighting strategy, which cuts down

the model parameters and drastically improves the computational efficiency. Wang et al. [20] proposed a multimodal sentiment analysis model that combines BERT-BiLSTM for text feature extraction and CNN with CBAM attention for reducing redundant information and enhancing cross-modal correlations. Setiadi et al. [21] proposed a BiGRU-BiDAF hybrid model for aspect-based sentiment analysis on e-commerce reviews, but the model is still incompetent in capturing long-term dependent information in the face of sequences of extreme lengths. Wang et al. [22] applied multi-level attention to adaptively fuse text and image modalities, improving sentiment representation learning. Xiao et al. [23] enhanced sentiment fusion by integrating attention mechanisms with graph convolutional networks, while Lin et al. [24] constructed unimodal and multimodal graphs to capture hierarchical relationships across modalities. Although these graph-based approaches have advanced the modeling of inter- and intra-modal relations, they primarily emphasize semantic interactions and contrastive structures among features, without explicitly accounting for modality-specific spatial or positional information.

In recent years, the field of natural language processing (NLP) has witnessed progress in addressing long-term dependency issues through Transformer models, particularly those employing self-attention mechanisms. Kim and Park [25] introduced a single-stream Transformer that alleviates the loss of modal characteristics during fusion and improves information integration. However, it assumes equal contributions from all modalities and lacks explicit structural modeling for non-text modalities. Recent vision-language pre-training (VLP) models, such as VL-BERT [26] and LXMERT [27], demonstrated strong visual-text alignment through unified or dual-encoder frameworks, while Zhang et al. [28] integrated BERT with ResNet50 for multimodal sentiment tasks. Although effective in mitigating inter-modal gaps, these approaches flatten or pool feature maps, which discards local dependencies in video frames and frequency–time structures in speech. Other strategies, such as Chandrasekaran et al. [29] combining LSTM with XGBoost, Cai et al. [30] employing multi-task fusion with attention, and PEST [31] for video sentiment analysis, aim to enhance cross-modal interaction but still treat non-text features as sequences without preserving internal spatial layouts. Similarly, Rahmani et al. [32] and Wang et al. [33] focus on cognitive cues and hierarchical fusion but neglect explicit positional encoding for modality-specific structures. Sun et al. [34] use a Transformer-based cross-modal interaction but largely overlook spatial information preservation. While these works strengthen semantic alignment and inter-modal correlation, they generally lack mechanisms to model spatial or structural properties of video and speech. This limitation may lead to feature loss during fusion.

Despite the advancements in previous studies, many existing models either overlook the structural spatial features of non-text modalities or treat all modalities as equally important. As a result, challenges persist in preserving critical modality-specific features and effectively utilizing positional information during the fusion process. The primary distinction between our proposed approach and existing work is that we regard the text modality as fundamental, processing it after integrating speech and video modalities. Before fusion, we encode the spatial positions of the speech and video modalities to ensure comprehensive utilization of their internal structures, such as time–frequency patterns in speech and spatial layouts in video. This enables the model to retain modality-specific characteristics that are often lost in flattened or sequence-based processing. By combining spatial position encoding with a fusion embedding mechanism, our method ensures comprehensive utilization of modality-specific features, ultimately leading to more accurate and robust sentiment analysis outcomes.

3 The Proposed Method

The proposed method can be categorized into three components, as illustrated in Fig. 2. The first component is spatial position encoding, which encodes both the speech and video modalities before fusion embedding and subsequently generates their outputs. The second component is the fusion embedding

module, which takes the outputs from the spatial position encoding as input and processes them through a fusion gate before sending the result to AOBERT [25] for processing the text, speech, and video modalities. We chose AOBERT over other Transformer-based models because it preserves inter-modal dependencies through a single-stream Transformer and joint pre-training tasks. This aligns well with our approach's focus on text and leveraging context-aware fusion to enhance cross-modal structural information. The comparison with other Transformer-based models is shown in the Table 1. The final component is the sentiment classification module.

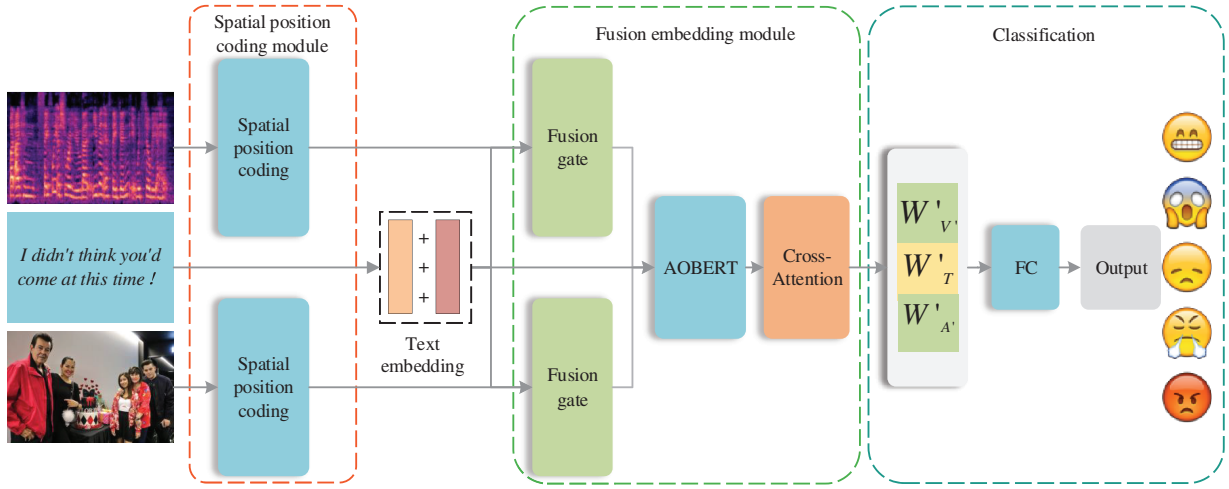

Figure 2: Overall framework

Table 1: The adaptability of representative Transformer-based models to text-centric multimodal strategies

Model	Adaptability
AOBERT [25]	Maintaining inter-modal dependencies through a single-stream Transformer and joint pre-training tasks aligns well with our approach.
LXMERT [27]	Cross-attention allows anchoring on text, but fusion still symmetric by default.
BERT-ResNet50 [28]	This approach flattens or pools feature maps, thereby discarding the spatial structure in video and speech.
ICCN [34]	Effective for semantic alignment but lacks explicit spatial/structural preservation, limiting text-centered extensions.

In this study, the three input modalities, text, speech, and video, are defined as: $X_T \in R^{d_T \times L}$, $X_A \in R^{d_A \times L}$, $X_V \in R^{d_V \times L}$. These modal vectors possess a fixed length L and dimensions d_T, d_A, and d_V. To ensure uniformity among inputs when some modalities are shorter than L, zero padding is employed. Our model also utilizes modal pairs defined as (X_T, F_A) and (X_T, F_V). Text serves as the base modality in the model, defined for the modal pairs as: $T = (X_T)$, $A' = (X_T, F_A)$, $V' = (X_T, F_V)$.

3.1 Spatial Position Encoding

Due to the complexity of multimodal data structures, speech and video modalities are prone to information loss during encoding. To preserve their intrinsic spatial and temporal characteristics, we

introduce a 2D spatial position encoding scheme. Unlike the standard 1D positional encodings used in NLP tasks or fixed 2D encodings in vision models, our method applies row-wise and column-wise sinusoidal encodings to the feature matrices of each modality, explicitly preserving frequency-time relations in speech and spatial layouts in video. Inspired by the sinusoidal positional encoding in the Transformer model [13], we extend it to encode two independent axes—row and column—denoted as PE_R and PE_C in Eqs. (1) and (2). These encodings are computed separately and then combined (e.g., via addition or concatenation) to produce the final position-aware representation. For implementation, input features are organized into matrices of size $d \times L$, where d and L represent rows and columns, respectively—spatial height and width for video, or frequency bins and time frames for audio. This structure-aware encoding strategy enhances modality-specific representation and supports effective downstream fusion, as illustrated in Fig. 3.

$$PE_R = \begin{cases} \sin\left(\frac{k}{10000^{\frac{2q}{d_{model}}}}\right), \\ \cos\left(\frac{k}{10000^{\frac{2q}{d_{model}}}}\right), \end{cases} \quad k \in [1, 40], \tag{1}$$

$$PE_C = \begin{cases} \sin\left(\frac{j}{10000^{\frac{2q}{d_{model}}}}\right), \\ \cos\left(\frac{j}{10000^{\frac{2q}{d_{model}}}}\right), \end{cases} \quad j \in [1, 768], \tag{2}$$

where k and j represent the position of the token in the sequence, q and $2q$ denote the odd and even dimensions, respectively, d_{model} denotes that the dimension of the token is 512 dimensions.

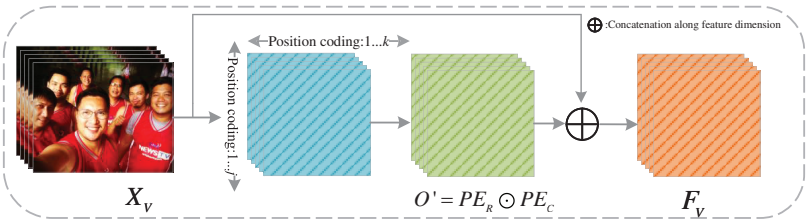

Figure 3: Spatial position encoding

Before applying the encoding, we define the row and column position embedding vectors as R and C, which represent positional information along the frequency/height and time/width axes, respectively. To ensure the completeness of the positional information, we adhere to a specific correspondence rule: odd positions correspond to odd positional information, while even positions correspond to even positional information. After spatial position encoding, the row position vectors, column position vectors, and the encoded matrix are O' defined as described in Eqs. (3) and (4).

$$R = [R_0, R_1, R_2, \ldots, R_k]^T, C = [C_0, C_1, C_2, \ldots, C_j]^T, \tag{3}$$

$$O' = R \odot C^T. \tag{4}$$

After spatial positional encoding, we combine the encoded features with the original data via a weighted summation to retain key information while minimizing feature loss. To balance their contributions, we introduce a weight parameter α, enabling flexible integration that preserves the original modality's characteristics

while leveraging the structural benefits of spatial encoding. The final output is shown in Eq. (5).

$$F = (1-\alpha)\cdot O + \alpha \cdot O', F \in \{V, A\}, \tag{5}$$

where O and α represent the original information and the weight parameter, respectively.

3.2 Fusion Embedding

The fusion embedding module aims to address the limitations of traditional fusion strategies, such as tensor-based methods that suffer from dimensionality explosion and attention-based fusion methods that assume equal modality contributions. Our design integrates textual embeddings with a lightweight Fusion Gate, which performs dimension alignment and adaptive feature integration in a computationally efficient manner.

The textual embedding component consists of token encoding and positional encoding. Token encoding converts individual words from the text X_T into numerical representations, while positional encoding provides essential positional information for the textual sequence. The text embedding employs the standard BERT embedding method; the outputs of the text embedding and spatial position encoding are collectively used as inputs for the fusion process. Unlike simple concatenation, the Fusion Gate applies a linear transformation and LayerNorm to stabilize scale differences while adaptively weighting auxiliary modalities to prevent feature dominance.

After fusion through the Fusion Gate, the resulting representations (A' and V') are combined with the text stream to form three inputs. To further model deep contextual interactions across modalities, we employ an adapted AOBERT as the backbone for multimodal reasoning. Unlike the original single-stream design, AOBERT in our framework processes three streams—text, text-audio, and text-video—enabling hierarchical refinement while keeping text as the core modality. This architecture enables our proposed Spatial Position Encoding and Fusion Embedding to be naturally integrated into the model, while AOBERT's layer-by-layer fusion strategy provides a flexible structure for dynamic cross-modal interactions. Cross-attention layers are leveraged for fine-grained integration, and the Pooler layer, followed by a fully connected layer with tanh activation, produces task-specific sentiment representations. The overall interaction flow is illustrated in Fig. 4.

$$F' = Linear(F), \tag{6}$$

$$T \oplus F = ([A, V]), \tag{7}$$

$$X_{A'} = LN(X_T \oplus F_A), X_{V'} = LN(X_T \oplus F_V), \tag{8}$$

where T represents a text modality, F denotes an speech modality or a video modality, $\oplus$ represents concatenation along the feature, dimension $T \oplus F$ is defined as the fusion of T and F.

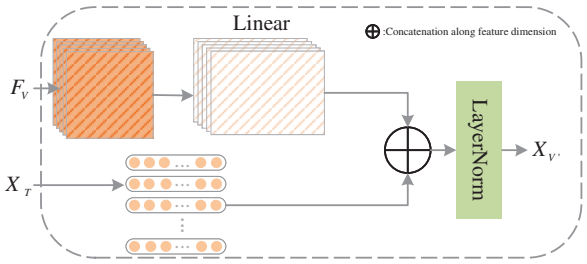

Figure 4: Fusion gate

Due to the distinct processing methods employed for these modalities, their dimensions differ. We establish the text modality as the baseline, enabling the sizes of the other two modalities to adjust accordingly to align with the text modality's dimensions. The specific procedure is as follows: F is projected through the linear layer in Eq. (6) to connect with the textual modality A. Subsequently, T and F are combined based on the sequence length operation outlined in Eq. (7).

To mitigate the scale discrepancies between modalities and enhance model stability, we implement a LayerNorm layer with a regularized dimension of d_T. Consequently, the dimensions and lengths of $X_{A'}$ and $X_{V'}$, which are processed by the fusion gate in Eq. (8), are both $R_T \times 2L$, while the length of the text modality T remains L. Downstream tasks are conducted using the Pooler layer in AOBERT, which incorporates a fully connected layer alongside the tanh activation function. When the three pairs T, A', and V' are processed by the Pooler layer, the following output is shown in Eq. (9). The detailed forward computation of the Fusion embedding is summarized in Algorithm 1.

$$W = \tanh(Linear(Z)), Z \in \{T, A', V'\}. \tag{9}$$

Algorithm 1: Forward computation of the Fusion embedding

Input: Text embedding $X_T \in \mathbb{R}^{d_T \times L}$, auxiliary modalities $\mathcal{F} = \{F_A, F_V\}$ where $F \in \mathbb{R}^{d_F \times L}$
Output: Fused streams $\{X_T, X_{TA}, X_{TV}\}$, with $X_T \in \mathbb{R}^{d_T \times L}$ and $X_{T*} \in \mathbb{R}^{d_T \times 2L}$
1: Initialize outputs: $X_T \leftarrow X_T$
2: **for** each auxiliary modality F in $\mathcal{F}$ **do**
3: $F' \leftarrow \text{Linear}_{\text{proj}}(F)$
4: $Z \leftarrow X_T \oplus F'$ // concatenate along **length** axis $\rightarrow Z \in \mathbb{R}^{d_T \times 2L}$
5: $Z' \leftarrow \text{LayerNorm}(Z)$
6: $X_{TF} \leftarrow \text{AOBERT\_stream}(Z')$ // cross-attention implicitly learns modality weighting
7: Store X_{TF} as X_{TA} or X_{TV} accordingly
8: **end for**
9: **return** $\{X_T, X_{TA}, X_{TV}\}$

3.3 Sentiment Classification

To achieve meaningful fusion performance, we introduce a cross-attention mechanism to process W_T, W_A, W_V. Fig. 5 provides a detailed illustration of the cross-attention layer's structure, including an example of how W_T is processed. After passing through the cross-attention layer, the transformed representations W'_T, $W'_{A'}$ and $W'_{V'}$ are obtained. Subsequently, each set of multimodal fusion representations undergoes a connectivity operation, and the results are predicted using a fully connected layer.

The fusion loss $Loss_f$ is computed across the three pairs of modal data, yielding three types of $Loss_T$, $Loss_{A'}$, and $Loss_{V'}$. This methodology is similarly applied to compute the loss values for each type. Given that $Loss_{task}$ is task-related and sentiment analysis is fundamentally a regression problem, we employ a mean squared error loss function. Ultimately, the model is optimized by minimizing the total average of the three fusion losses, which are influenced by both $Loss_{task}$ and $Loss_f$ (as defined in Eqs. (10) and (11)), thereby enhancing its performance across these combined objectives.

$$Loss_{task} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2, \tag{10}$$

$$Loss = Loss_{task} + \frac{Loss_T + Loss_{A'} + Loss_{V'}}{3}. \tag{11}$$

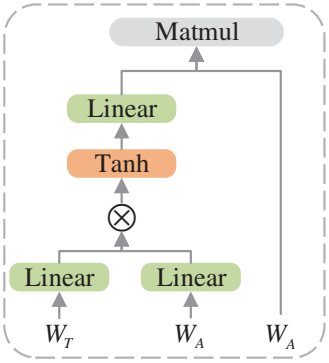

Figure 5: Self-attention layer in a classifier

4 Experiments

In this section, we conduct a series of experimental evaluations to assess the efficacy of our novel model and benchmark its performance against alternative models across the two datasets under evaluation. All experimental codes are available on GitHub[1].

4.1 Experimental Datasets

We conduct our experiments utilizing two publicly available datasets, CMU-MOSI [35] and CMU-MOSEI [36], each of which is comprehensively outlined in the following.

CMU-MOSI: A comprehensive multimodal corpus of emotional intensity, curated and structured by Zadeh et al. [35], draws upon YouTube content where speakers articulate their perspectives on film themes. This extensive resource encompasses a diverse sample of 2199 video clips featuring English monologues from individuals of various ethnic backgrounds. The emotional intensity of each video within this corpus spans a range from −3 to +3. The dataset has been split into training, validation, and testing sets with respective sample counts of 1284, 229, and 686, as illustrated in Table 2. This partitioning enables us to train the model, validate its performance, and evaluate its generalization ability.

Table 2: Dataset details

Dataset	CMU-MOSI	CMU-MOSEI
Training set	1284	16,216
Validation set	229	1871
Testing set	686	4654

CMU-MOSEI is an extension of the CMU-MOSEI dataset by Zadeh et al. [36]. It consists of 23,453 annotated video clips. This corpus contains both affective and emotional labels. Emotion intensity is annotated using the same method as CMU-MOSI. The dataset encompasses a comprehensive spectrum of

[1] https://github.com/WahPr/SPEFE (accessed on 25 August 2025)

emotions, spanning from negative to positive across seven distinct categories. The intensity of these emotions is quantitatively represented on a scale ranging from −3 to +3. The dataset classification is detailed in Table 2.

We process the CMU-MOSI and CMU-MOSEI sentiment datasets using the CMU-Multi-modal Data SDK [37] toolkit.

Text feature extraction: In previous studies, GloVe word embeddings [38] were utilized as textual modal features for each marker. Currently, the most advanced research findings have been achieved by employing pre-trained BERT as a feature extractor for textual discourse.

Video feature extraction: Both CMU-MOSI and CMU-MOSEI leveraged the Facet tool to characterize facial attributes, employing the Facial Action Coding System (FACS) methodology for extracting fundamental and intricate emotional nuances present within each video frame. The dimensions of the video features extracted by the toolkit are 47 and 35, respectively.

Speech feature extraction: The speech features in the dataset were partially processed using the COVAREP [39] algorithm, a dedicated speech signal processing algorithm designed to derive low-level speech attributes. This processing resulted in 74-dimensional speech feature vectors.

For all experiments, we randomly split each dataset into training, validation, and test sets from the same source with identical distributions. A fixed random seed was used to ensure that the same data partition was consistently applied to all models under comparison, thereby eliminating the influence of sampling variability. Furthermore, to reduce the impact of stochastic factors such as parameter initialization and optimizer dynamics, we repeated each experiment five times and report the average performance across runs.

4.2 Comparison Models

We select a series of models for comparative analysis that include metrics for evaluating sentiment analysis, aiming to demonstrate their excellent performance and advantages in addressing multimodal sentiment analysis through rigorous experiments. These models tackle challenges in multimodal sentiment analysis from different perspectives: they address the forgetting problem using long-term and short-term memory networks [37,40–42], facilitate multimodal data fusion utilizing tensor fusion [10,19], and enhance the interaction between multimodal information by employing transformers [14,43]. The models used for comparison are summarized in Table 3. All models are tested on the same dataset, ensuring the fairness and comparability of the experimental results.

Table 3: Representative multimodal models

Model	Description
LMF [10]	A low-rank multimodal fusion method.
MulT [14]	A multimodal converter-based model for sentiment analysis.
TFN [19]	A network model that fuses tensors through a specific architecture.
Graph-MFN [36]	A memory fusion network with an advanced gating mechanism.
BC-LSTM [40]	A bi-directional contextual LSTM model.
MFM [41]	A multimodal decomposition model for sentiment analysis.
MFN [42]	A model that uses gated memory fusion for multimodal integration.
MCR [43]	A co-enhancement technique for target and source modes to aid cross-modal fusion.
MCTN [44]	A model that learns robust joint representations through transitions between modes.

(Continued)

Table 3 (continued)

Model	Description
ICCN [34]	A multimodal sentiment analysis model using Transformer architecture.
GraphCAGE [45]	An unaligned multimodal sequence model based on GNN and capsule networks.
EF-HEMT [46]	A holographic representation approach for higher-order multimodal fusion.
LMR-CBT [47]	A CB-Transformer-based model for learning from unaligned multimodal sequences.
MAG-BERT [48]	A BERT-based model with multimodal adaptation gating.
MISA [49]	A model learning both modality-invariant and modality-specific representations.

4.3 Evaluation Metrics and Parameter Settings

Sentiment analysis is fundamentally regarded as a regression task that aims to predict the strength or polarity of emotional tendencies. Consequently, we select Mean Absolute Error (*MAE*) as one of the primary performance indicators, supplemented by the correlation coefficient (*Corr*) to assess the strength of the linear relationship between predicted and actual values. However, it is important to note that sentiment analysis often manifests as a classification task, wherein the emotional tendency of a text is classified as positive, negative, or neutral. Accordingly, we also incorporate Accuracy for Binary Classification (*Acc*), which directly measures the correctness of classification results, and combine it with the F1 Score (*F1*), the harmonic mean of Precision and Recall, to provide a comprehensive evaluation of classification performance, as defined in Eqs. (12) and (13).

$$Acc = \frac{TP + TN}{TP + TN + FP + FN}, \tag{12}$$

$$F1 = \frac{2TP}{2TP + FP + FN}, \tag{13}$$

where True Positive (*TP*), True Negative (*TN*), False Positive (*FP*), and False Negative (*FN*) include correct positive predictions, correct negative predictions, incorrect negative predictions, and incorrect positive predictions, respectively. These categories collectively serve as essential metrics for assessing the classification accuracy and performance of the model.

In this research, we utilize a server equipped with 24 GB of NVIDIA RTX 3090 graphics processing unit (GPU), employing Python 3.8 as the programming language and PyTorch 1.10 as the deep learning framework. For our experiments, we leverage the pre-trained BERT base uncased model to extract the word embedding matrix. During backpropagation, we employ the AdamW optimizer, which incorporates a linear learning rate warm-up strategy. The learning rate is set to 5e-4, with a weight decay of 0.1 and a dropout rate of 0.38. The training regimen involves a batch size of 16 and a total of 100 iterations. For the MOSI and MOSEI datasets, the input length is fixed at 40 tokens. The optimal experimental parameter settings we obtained are shown in Table 4.

Table 4: Parameter settings

Parameter	Value
Batch size	16
Learning rate	5e−4
Dropout	0.38
Weight decay	0.1
Max length	40
Epoch	100

4.4 Results and Discussions

4.4.1 Comparison on CMU-MOSI Dataset

The results on the CMU-MOSI dataset are shown in Table 5. Our model achieves competitive performance, outperforming several representative baselines such as MFN [42], MCTN [44], and BC-LSTM [40] across accuracy, F1-score, and correlation metrics. This demonstrates the effectiveness of our structured fusion strategy and spatial position encoding in preserving modality-specific features.

Table 5: Comparison on CMU-MOSI dataset

Model	*Acc* ↑	*MAE* ↓	*Corr* ↑	*F1* ↑
BC-LSTM [40]	73.9	1.079	0.581	73.9
TFN [19]	73.9	0.970	0.633	73.4
LMF [10]	76.4	0.912	0.668	75.7
MFN [42]	77.4	0.965	0.632	77.3
MCTN [44]	79.1	0.909	0.676	79.1
MFM [41]	78.1	0.951	0.662	78.1
MulT [14]	83.0	0.871	0.698	82.8
ICCN [34]	83.0	0.860	0.710	83.0
GraphCAGE [45]	82.1	0.933	0.684	82.1
EF-HEMT [46]	82.3	0.901	0.701	82.5
LMR-CBT [47]	81.2	–	–	81.0
MAG-BERT [48]	84.2	0.712	0.796	84.1
MISA [49]	81.8	0.783	0.761	81.7
Our method	81.7	0.896	0.657	81.6

While recent transformer-based models like MAG-BERT [48] and MulT [14] achieve higher accuracy, they rely on large-scale pre-trained models or complex temporal alignment mechanisms. Our approach, in contrast, focuses on spatial modeling and modular fusion, offering better interpretability and structural flexibility. It is worth noting that MISA [49] also performs well by learning modality-invariant and modality-specific representations. Compared to it, our model maintains competitive classification results with a simpler and more modular design.

Although our model does not surpass the state-of-the-art ICCN model according to evaluation metrics, we attribute this limitation primarily to the constrained size of the MOSI dataset, which restricts our model's ability to fully leverage its potential during training. Overall, data augmentation addresses the challenge of

inadequate training data by artificially increasing the dataset's size and diversity. While various strategies exist for augmenting unimodal data, there is currently a lack of tailored techniques specifically designed to tackle the complexities of multimodal data.

Fig. 6 shows the confusion matrix results of the model on the CMU-MOSI dataset. It can be observed from this matrix that the model performs well in the prediction of most samples and is capable of accurately identifying and predicting their correct categories. However, it should also be noted that there are still a small number of samples whose prediction results have deviated and failed to be classified correctly. Fig. 7 presents the visualization diagram of the model after dimensionality reduction processing. It can be intuitively seen in this figure that samples belonging to the same category can cluster well together in the feature space, demonstrating the effectiveness of the model in classification tasks. However, at the same time, there is also a situation where the two types of samples are difficult to completely separate in the feature space. This might be due to the relatively small number of samples in the dataset, resulting in the model failing to fully learn the subtle differences between various categories during the training process.

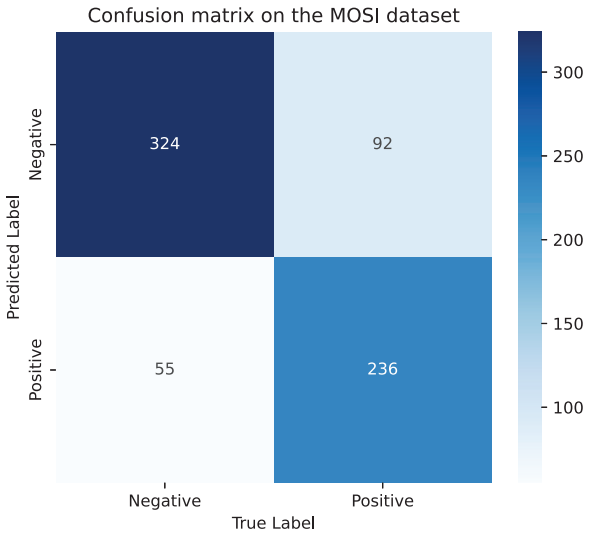

Figure 6: Confusion matrix on the CMU-MOSI dataset

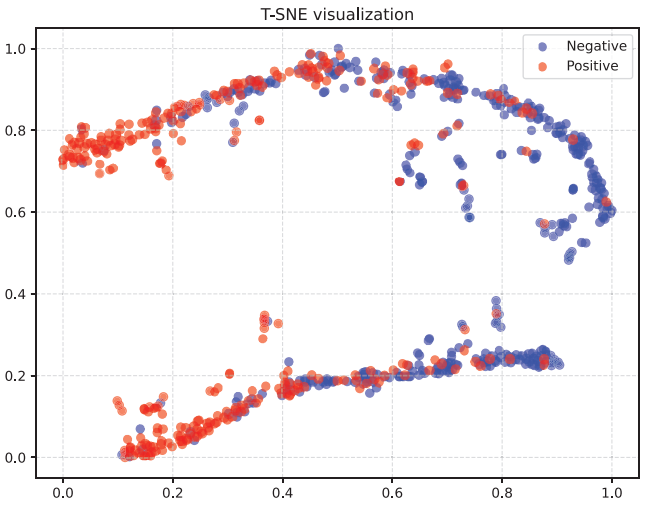

Figure 7: Visualization results on the CMU-MOSI dataset. The x- and y-axes are 2D embedding coordinates without physical meaning, used for visualizing sample relationships

4.4.2 Comparison on CMU-MOSEI Dataset

The results of sentiment analysis performed on the CMU-MOSEI dataset are depicted in Table 6. The proposed model shows improved performance, yielding observable improvements across various classification and regression metrics. When compared to existing models, our model exhibits comparable or superior performance, particularly excelling in *Acc*, *Corr*, and *F*1 metrics. This indicates the model's strength in leveraging emotional features, capturing and utilizing emotional information from the data.

Table 6: Comparison on CMU-MOSEI dataset

Model	*Acc* ↑	*MAE* ↓	*Corr* ↑	*F1* ↑
Graph-MFN [36]	76.9	0.710	0.540	77.0
MCTN [34]	79.8	0.609	0.670	80.6
TFN [19]	82.5	0.593	0.700	82.1
LMF [10]	82.0	0.623	0.677	82.1
MFM [41]	84.4	0.568	0.717	84.3
MulT [14]	82.5	0.580	0.703	82.3
ICCN [34]	84.2	0.565	0.713	84.2
MCR [43]	84.7	0.554	0.736	84.3
GraphCAGE [45]	81.7	0.609	0.670	81.8
EF-HEMT [46]	81.9	0.597	0.699	82.2
LMR-CBT [47]	80.9	–	–	81.5
MAG-BERT [48]	84.7	–	–	84.5
MISA [49]	83.6	0.555	0.756	83.8
Our method	85.8	0.569	0.759	85.5

As shown in Table 6, our model achieves the best overall performance on the CMU-MOSEI dataset, outperforming recent methods including MAG-BERT [48] and MISA [49]. While MAG-BERT leverages large-scale pre-trained transformers with modality adaptation gates to enhance language representations, it increases model complexity and training cost. MISA employs a dual-branch strategy to disentangle modality-invariant and modality-specific features. In comparison to these and earlier fusion techniques such as Tensor Fusion Network (TFN) [19] and Low-rank Multimodal Fusion (LMF) [10], our framework explicitly integrates spatial modeling, contributing to superior accuracy, correlation, and F1 scores. Other competitive models like the Multimodal Factorization Machine (MFM) [41] and Multimodal Contextual Reinforcement (MCR) [43] enhance multimodal interactions; MFM adeptly decomposes features into discriminative and modality-specific components, and MCR creatively integrates novel target-modal reinforcement mechanisms to robustly enhance representations and capture complex interactions. Despite these sophisticated strategies, our approach consistently surpasses them, validating the advantages of our modular design and delivering better performance with a more compact and interpretable architecture.

In contrast, our model achieves enhancements by optimizing the fusion process and encoding the spatial positions of speech and video modalities. This optimization allows for more accurate data fusion from various modalities and utilizes spatial position information to enhance performance. The proposed architecture introduces additional computation primarily in the fusion gate and positional encoding stages, but the increase remains moderate compared to tensor-based methods, ensuring practical feasibility. However, despite these improvements, the model may still face limitations in certain conditions due to its underlying

assumptions. Specifically, performance may degrade under severe modality asynchrony or extremely noisy audio/video inputs, as the positional encoding assumes relatively stable structural patterns.

Notably, the comparative models, including TFN, MFN, MFM, and ICCN, demonstrated improved performance on the MOSEI dataset relative to the CMU-MOSI dataset. This enhancement can be attributed to the richer and more diverse data samples in MOSEI, which enabled these models to undergo a more comprehensive training process and realize their full potential. This observation highlights the crucial importance of increasing dataset size to improve model performance and underscores the impact of data richness on refining model capabilities.

Fig. 8 shows the confusion matrix results on the CMU-MOSEI dataset. It can be seen that the model can predict the correct categories of most samples, but there are still a small number of samples with incorrect predictions. Compared with the results of the CMU-MOSI dataset, the prediction accuracy rate has been improved. Fig. 9 is the visualization diagram of the model after dimensionality reduction. From the figure, it can be intuitively observed that samples of the same category can be closely clustered together to form a distinct cluster structure, and at the same time, the discrimination between samples of different categories is also very significant. Compared with the visualization results on the CMU-MOSI dataset, the model performs better on the CMU-MOSEI dataset, which further confirms the generalization ability of the model on different datasets. However, it is worth noting that although the overall performance of the model is excellent, there are still some samples that are difficult to clearly distinguish in the visualization graph, thereby affecting the performance of the model.

4.5 Ablation Study

In this section, we conduct a comprehensive series of ablation studies to investigate the specific contributions of the spatial position encoding component, with the results summarized in Table 7.

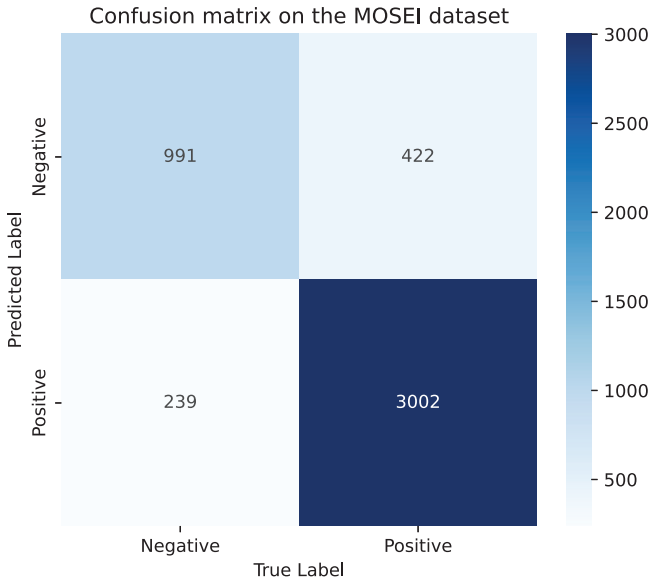

Figure 8: Confusion matrix on the CMU-MOSEI dataset

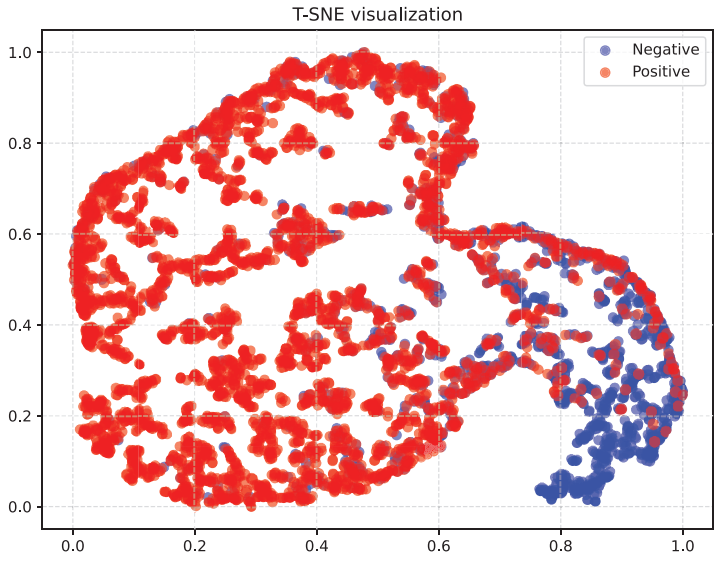

Figure 9: Visualization results on the CMU-MOSEI dataset. The x- and y-axes are 2D embedding coordinates without physical meaning, used for visualizing sample relationships

Table 7: Ablation results encoded by spatial position

Dataset	*Acc* ↑	*MAE* ↓	*Corr* ↑	*F1* ↑
CMU-MOSI	81.7	0.896	0.657	81.6
CMU-MOSI (ablation)	80.1	0.928	0.653	80.2
CMU-MOSEI	85.8	0.559	0.759	85.5
CMU-MOSEI (ablation)	83.6	0.556	0.753	83.7

Experimental results from the CMU-MOSI dataset indicate that the unablated model achieves improvements of 1.6% and 1.4% in *Acc* and *F1* metrics, respectively, compared to the ablated model. This level of improvement signifies a notable enhancement in model performance. Additionally, the unablated model demonstrates a reduction in mean absolute error (*MAE*) and an enhancement in correlation (*Corr*) metrics relative to the ablated model, providing further evidence of optimization in several performance aspects. This substantial performance enhancement is primarily attributed to our proposed spatial position encoding module. By encoding speech and video modalities, the model minimizes the loss of unique features from each modality, allowing it to retain and utilize critical information more effectively. Simultaneously, it optimizes the use of internal features of each modality, making the model more adept at handling complex multimodal data. Collectively, these advantages lead to improvements in overall model performance, particularly in *Acc* and *F1* metrics, further validating the effectiveness and sophistication of the spatial position encoding module.

Similarly, the fully trained model evaluated on the CMU-MOSEI dataset outperforms the ablated model, underscoring the importance of the spatial position encoding module. The unablated model achieves improvements of 2.2% and 1.8% in *Acc* and *F1* metrics, respectively, providing additional evidence of the model's validity. Moreover, the unablated model exhibits notable gains in the *Corr* metric, suggesting enhanced overall performance when managing multimodal data. However, it is important to acknowledge that the unablated model is not optimal concerning the *MAE* index. The broader spectrum of emotions

and diverse scenarios represented in the CMU-MOSEI dataset pose greater challenges for the model, contributing to the observed suboptimal performance in the *MAE* index. Nonetheless, the unablated model exhibits improved generalization ability on this dataset, captures more nuanced features, and shows enhancements across multiple metrics. These results further corroborate the effectiveness of the proposed spatial position encoding module, which minimizes the loss of modality-specific features and improves the utilization of intramodal features, thereby enhancing overall model performance. Although there remains potential for further improvement in the *MAE* metrics, the model's performance in other metrics clearly demonstrates its superiority and potential.

We also conducted ablation experiments of the fusion embedded module to verify the effectiveness of the module. The experimental results on the two data sets are shown in Table 8. The experimental results on the CMU-MOSI dataset show that compared with the ablation model, the non-ablation model achieves a increase of 2.1% in *Acc* metric and 1.9% in *F1* metric, respectively, which indicates that the model performance has been improved. At the same time, the *MAE* metric decreased and the *Corr* metric increased in the non-ablation model compared with the model with the fusion embedded module. The performance improvement of the model is mainly due to the fusion embedded module proposed by us. By integrating text as the basic mode with audio and video modes, the key features of each mode can be retained to the greatest extent, thus improving the overall performance of the model, which further verifies the effectiveness of the fusion embedded module.

Table 8: Ablation results of fusion embedding

Dataset	*Acc* ↑	*MAE* ↓	*Corr* ↑	*F1* ↑
CMU-MOSI	81.7	0.896	0.657	81.6
CMU-MOSI (ablation)	79.6	0.925	0.646	79.7
CMU-MOSEI	85.8	0.559	0.759	85.5
CMU-MOSEI (ablation)	83.1	0.572	0.745	83.4

Similarly, on the CMU-MOSEI dataset, the non-ablated model achieved a improvement of 2.7% in the *Acc* metric and 2.1% in the F1 metric compared with the ablated model, which further verified the importance of the proposed module in improving the overall performance of the model. At the same time, *MAE* metric decreased and *Corr* metric increased slightly, which further verified the advantage of the model in capturing the correlation between data. In summary, the experimental results on the two datasets consistently verify that our proposed fusion embedded module can improve the overall performance of the model.

Although both components show similar performance drops in isolation, the fusion embedding has a more direct impact on cross-modal integration. The spatial position encoding acts as a supporting module, and its effectiveness depends on whether the fusion mechanism can leverage the encoded structural cues. Without proper fusion, spatial features may not be fully utilized. This highlights the complementary nature of the two modules.

In the spatial position encoding module, we aim to maximize the utility of modal feature information by integrating it with the original information. This strategy enhances the features encoded by spatial position while preserving the original data, thus enriching the overall feature representation. However, given the potential discrepancies between the information processed by the spatial position encoding module and the original data, a straightforward addition may not represent the optimal fusion method. To achieve more precise control over the contributions from the two sources, we introduce a weighting parameter α. This parameter adjusts the balance between the spatially position-encoded modal information and the original

data in the final fused output, enabling more flexible and accurate integration. To determine the optimal value of α, we conducted a series of experiments with values ranging from 0.1 to 0.9 in increments of 0.1. These experiments were designed to identify the setting that yields the best model performance, thereby validating the effectiveness of the spatial position encoding module and exploring its most suitable role in multimodal information fusion.

Table 9 presents the experimental results for the CMU-MOSI dataset. Our findings indicate that the model's accuracy (*Acc*), mean absolute error (*MAE*), and F1 metrics reached their optimal levels when the weighting parameter α was set to 0.7, although the correlation metric (*Corr*) did not attain its highest value. As a relatively small dataset, CMU-MOSI may exhibit an imbalanced data distribution, which poses challenges for effective model training. This imbalance can cause the model to become biased toward the majority class, thereby compromising its ability to correctly classify the minority class. Notably, when $\alpha = 0.7$, the model demonstrates peak performance across the *Acc*, *MAE*, and *F*1 metrics. This configuration likely achieves a more effective balance in the loss function, enabling the model to better capture features that are critical for accurate classification. The improvement in *Acc* reflects enhanced classification accuracy; the reduction in *MAE* indicates lower errors in predicting sentiment intensity; and the increase in *F1* shows a more favorable trade-off between precision and recall. Additionally, the limited number of samples in CMU-MOSI may restrict the model's ability to fully learn representative features, thus constraining its generalization capacity.

Table 9: Experimental results of weight share on CMU-MOSI dataset

α	*Acc* ↑	*MAE* ↓	*Corr* ↑	*F1* ↑
0.1	79.1	0.928	0.652	78.9
0.2	79.2	0.905	0.658	79.1
0.3	78.1	0.947	0.639	78.1
0.4	79.6	0.911	0.649	79.5
0.5	78.8	0.908	0.653	78.7
0.6	78.9	0.912	0.657	78.8
0.7	81.7	0.896	0.657	81.6
0.8	79.1	0.906	0.661	78.8
0.9	78.6	0.905	0.662	78.5

Table 10 presents the experimental results on the CMU-MOSEI dataset. Through detailed analysis, we found that the model achieves optimal performance across all evaluation metrics—except for the *MAE*—when the weighting parameter α is set to 0.3. We hypothesize that this result may be attributed to the comprehensive nature of the CMU-MOSEI dataset, which contains abundant speech and video data. This richness provides the model with ample opportunities to learn expressive feature representations, thereby enhancing its performance across various sentiment analysis tasks. However, we also observe variations in the model's use of speech and video features under different α values. These variations may result from how different settings of α influence the model's preference during feature extraction and fusion, ultimately affecting the optimization of the *MAE* metric. The experimental results underscore the importance of carefully tuning the weighting ratio between the original and position-encoded modal information within the spatial position encoding module, as this adjustment impacts overall model performance. This finding highlights the crucial role of α in balancing raw and encoded data contributions during multimodal fusion.

Table 10: Experimental results of weight share on CMU-MOSEI dataset

α	*Acc* ↑	*MAE* ↓	*Corr* ↑	*F1* ↑
0.1	83.3	0.599	0.746	83.3
0.2	83.8	0.617	0.717	83.1
0.3	85.8	0.559	0.759	85.5
0.4	83.7	0.562	0.751	83.9
0.5	83.4	0.557	0.748	83.5
0.6	83.9	0.569	0.753	83.9
0.7	83.3	0.559	0.754	83.5
0.8	83.5	0.564	0.755	83.5
0.9	84.6	0.585	0.756	84.5

Tables 6 and 7 demonstrate that removing either the spatial position encoding or the fusion embedding consistently degrades performance across accuracy, F1, correlation, and MAE. The benefits arise not simply from added parameters. Spatial position encoding preserves the intrinsic 2D organization of audio (time-frequency patterns) and video (facial regions), reducing information loss and providing a structural prior that dampens frame-level noise. The fusion embedding module performs lightweight dimensional and scale alignment while anchoring integration on the text stream, preventing auxiliary modalities with high variance from overwhelming linguistic semantics and stabilizing subsequent attention. Their effects are complementary: spatial encoding provides structure, while fusion embedding enables effective integration. Similar degradation when each module is removed separately indicates neither subsumes the other.

Tables 8 and 9 further reveal that the weighting coefficient α in Eq. (5) balances raw features (O) and position-enhanced features (O'). On the smaller, less balanced MOSI dataset, a higher α (0.7) yields better classification by emphasizing the structured representation to reduce variance and mitigate overfitting to noise. Conversely, on the larger, more diverse MOSEI dataset, a lower α (0.3) is optimal, as abundant data allows learning useful local patterns directly from raw features; a high α would over-constrain subtle continuous variations crucial for correlation and regression. Minor metric mismatches indicate the optimal α for discrete classification vs. continuous metrics may differ slightly.

4.6 Error Analysis

To further illustrate the limitations of the proposed model, we analyze a representative misclassified case from the CMU-MOSEI dataset, shown in Fig. 10. The text in this sample reads: "But, if you're interested in owning this, you can find it on amazon dot com."

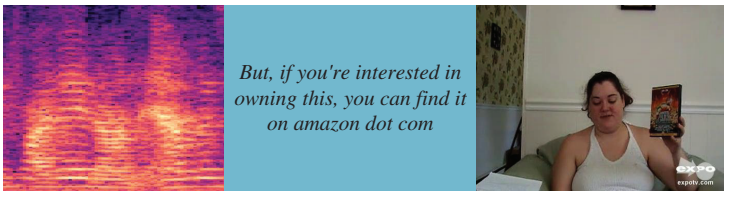

Figure 10: Misclassification cases (example from the CMU-MOSEI dataset)

The ground-truth sentiment label for this utterance is negative, yet the model incorrectly classified it as positive. This error can be primarily attributed to the dominance of text-based semantic cues in the fused

representation. Specifically, the sentence contains superficially positive lexical items such as "interested" and "owning", which are typically associated with favorable attitudes. In the absence of explicit negative sentiment words or strong negation patterns, the textual stream presents a semantic bias toward positivity.

Although our model integrates speech and visual modalities, in this case the prosodic and facial cues conveying disinterest or sarcasm were relatively subtle. The spatial position encoding mechanism preserved structural features from these modalities, but the fusion process may have assigned insufficient weight to them due to the stronger polarity signal inferred from the text. When auxiliary modalities carry low-intensity emotional cues, the model's decision boundary may still be disproportionately influenced by lexical sentiment priors from the text.

5 Conclusions and Future Work

In this paper, we proposed a multimodal sentiment analysis model that integrates fusion embedding and spatial position encoding. The model treats text as the primary modality and combines it with speech and video in a structured manner, preserving key modality-specific features. By applying spatial position encoding to speech and video, we mitigate spatial information loss and enhance feature utilization. Experiments on CMU-MOSI and CMU-MOSEI demonstrate the model performance, and ablation studies verify the contribution of each component.

In addition, our current experiments rely on benchmark datasets (CMU-MOSI and CMU-MOSEI) that provide relatively well-segmented and temporally synchronized text, audio, and video streams. However, real-world deployments often face modality asynchrony arising from heterogeneous sampling rates, imperfect speaker segmentation, frame dropping, network latency, and gradual drift between transcription timestamps and acoustic/visual signals. Such misalignment may weaken cross-modal attention and partially offset the benefits of spatial position encoding when temporal correspondence is ambiguous. Similarly, practical scenarios introduce diverse noise sources: background speech and music, reverberation, sensor or compression artifacts, facial occlusion, pose changes, tracking loss, and transcription or ASR errors. These factors can inject high-variance perturbations into time–frequency patterns or facial dynamics, potentially amplifying spurious correlations.

In future work, we will explore multi-task learning schemes, such as combining sentiment classification with intensity prediction, to further enhance model generalization. We also plan to investigate assigning primary modality roles dynamically, as well as integrating cross-modal attention mechanisms to adaptively weigh each modality. Moreover, we aim to incorporate interpretability techniques (e.g., Grad-CAM, attention visualization) to analyze the decision process, and address challenges related to temporal asynchrony in multimodal data. We also plan to perform fine-grained ablation studies on the fusion module by separately evaluating the effects of linear projection, layer normalization, and gating operations. To broaden applicability, we will also extend our spatial encoding strategy to non-grid modalities such as raw text via learnable or adaptive position encoding methods. Finally, we plan to include statistical tests and clustering metrics to strengthen the reliability of model evaluations.

We hope this modular framework serves as a foundation for future research in structured and interpretable multimodal fusion, facilitating progress in tasks beyond sentiment analysis.

Acknowledgement: Not applicable.

Funding Statement: This work was supported by the Collaborative Tackling Project of the Yangtze River Delta Sci-Tech Innovation Community (Nos. 2024CSJGG01503, 2024CSJGG01500), Guangxi Key Research and Development Program (No. AB24010317), and Jiangxi Provincial Key Laboratory of Electronic Data Control and Forensics (Jiangxi Police College) (No. 2025JXJYKFJJ002).

Author Contributions: The authors confirm contribution to the paper as follows: Chenquan Gan: Conceptualization, Validation, Writing—original draft, Writing—review & editing; Xu Liu, Yu Tang: Methodology, Writing—original draft, Formal analysis; Xianrong Yu, Qingyi Zhu, Deepak Kumar Jain: Supervision. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: All datasets used in this study are publicly available, and all referenced methods are from peer-reviewed, accessible publications.

Ethics Approval: Not applicable.

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

References

1. Mao Y, Liu Q, Zhang Y. Sentiment analysis methods, applications, and challenges: a systematic literature review. J King Saud Univ Comput Inf Sci. 2024;36(4):102048. doi:10.1016/j.jksuci.2024.102048.
2. Li Z, Guo Q, Pan Y, Ding W, Yu J, Zhang Y, et al. Multi-level correlation mining framework with self-supervised label generation for multimodal sentiment analysis. Inf Fusion. 2023;99(6):101891. doi:10.1016/j.inffus.2023.101891.
3. Zhao S, Jia G, Yang J, Ding G, Keutzer K. Emotion recognition from multiple modalities: fundamentals and methodologies. IEEE Signal Process Mag. 2021;38(6):59–73. doi:10.1109/MSP.2021.3106895.
4. Guo Z, Yu K, Li Y, Srivastava G, Lin JC. Deep learning-embedded social Internet of Things for ambiguity-aware social recommendations. IEEE Trans Netw Sci Eng. 2022;9(3):1067–81. doi:10.1109/TNSE.2021.3049262.
5. Chandrasekaran G, Nguyen TN, Hemanth DJ. Multimodal sentimental analysis for social media applications: a comprehensive review. Wires Data Min Knowl Discov. 2021;11(5):e1415. doi:10.1002/widm.1415.
6. Lai S, Hu X, Xu H, Ren Z, Liu Z. Multimodal sentiment analysis: a survey. Displays. 2023;80(2):102563. doi:10.1016/j.displa.2023.102563.
7. Ghosal D, Akhtar MS, Chauhan D, Poria S, Ekbal A, Bhattacharyya P. Contextual inter-modal attention for multi-modal sentiment analysis. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Brussels, Belgium. Stroudsburg, PA, USA: ACL; 2018. p. 3454–66. doi:10.18653/v1/d18-1382.
8. Baltrusaitis T, Ahuja C, Morency LP. Multimodal machine learning: a survey and taxonomy. IEEE Trans Pattern Anal Mach Intell. 2019;41(2):423–43. doi:10.1109/TPAMI.2018.2798607.
9. Gandhi A, Adhvaryu K, Poria S, Cambria E, Hussain A. Multimodal sentiment analysis: a systematic review of history, datasets, multimodal fusion methods, applications, challenges and future directions. Inf Fusion. 2023;91(3):424–44. doi:10.1016/j.inffus.2022.09.025.
10. Liu Z, Shen Y, Lakshminarasimhan VB, Liang PP, Zadeh A, Morency LP. Efficient low-rank multimodal fusion with modality-specific factors. arXiv:1806.00064. 2018.
11. Wllmer M, Kaiser M, Eyben F, Schuller B, Rigoll G. LSTM-Modeling of continuous emotions in an audiovisual affect recognition framework. Image Vis Comput. 2013;31(2):153–63. doi:10.1016/j.imavis.2012.03.001.
12. Mai S, Sun Y, Zeng Y, Hu H. Excavating multimodal correlation for representation learning. Inf Fusion. 2023;91(2):542–55. doi:10.1016/j.inffus.2022.11.003.
13. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. In: Advances in neural information processing systems. Long Beach, CA, USA: Curran Associates Inc.; 2017. 30 p.
14. Tsai YH, Bai S, Liang PP, Kolter JZ, Morency LP, Salakhutdinov R. Multimodal transformer for unaligned multimodal language sequences. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics; 2019 Jul 28–Aug 2; Florence, Italy. Stroudsburg, PA, USA: ACL; 2019. p. 6558–69. doi:10.18653/v1/p19-1656.
15. Kumar A, Vepa J. Gated mechanism for attention based multi modal sentiment analysis. In: ICASSP 2020—2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); 2020 May 4–8; Barcelona, Spain: IEEE; 2020. p. 4477–81. doi:10.1109/ICASSP40776.2020.9053012.
16. Yang X, Feng S, Wang D, Zhang Y. Image-text multimodal emotion classification via multi-view attentional network. IEEE Trans Multimed. 2021;23:4014–26. doi:10.1109/TMM.2020.3035277.

17. Swain M, Routray A, Kabisatpathy P. Databases, features and classifiers for speech emotion recognition: a review. Int J Speech Technol. 2018;21(1):93–120. doi:10.1007/s10772-018-9491-z.
18. Wang J, Mou L, Ma L, Huang T, Gao W. AMSA: adaptive multimodal learning for sentiment analysis. ACM Trans Multimed Comput Commun Appl. 2023;19(3s):1–21. doi:10.1145/3572915.
19. Zadeh A, Chen M, Poria S, Cambria E, Morency LP. Tensor fusion network for multimodal sentiment analysis. arXiv:1707.07250. 2017.
20. Wang H, Li X, Ren Z, Yang D, Ma C. Exploring multimodal sentiment analysis via CBAM attention and double-layer BiLSTM architecture. arXiv:2303.14708. 2023.
21. Setiadi DRIM, Warto W, Muslikh AR, Nugroho K, Safriandono AN. Aspect-based sentiment analysis on E-commerce reviews using BiGRU and bi-directional attention flow. J Comput Theor Appl. 2025;2(4):470–80. doi:10.62411/jcta.12376.
22. Wang H, Ren C, Yu Z. Multimodal sentiment analysis based on multiple attention. Eng Appl Artif Intell. 2025;140(2):109731. doi:10.1016/j.engappai.2024.109731.
23. Xiao L, Wu X, Wu W, Yang J, He L. Multi-channel attentive graph convolutional network with sentiment fusion for multimodal sentiment analysis. In: ICASSP 2022—2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); 2022 May 23–27; Singapore: IEEE; 2022. p. 4578–82. doi:10.1109/ICASSP43922.2022.9747542.
24. Lin Z, Liang B, Long Y, Dang YX, Yang M, Zhang M, et al. Modeling intra-and inter-modal relations: hierarchical graph contrastive learning for multimodal sentiment analysis. In: Proceedings of the 29th International Conference on Computational Linguistics; 2022 Oct 12–17; Gyeongju, Republic of Korea: Association for Computational Linguistics; 2022. p. 7124–35.
25. Kim K, Park S. AOBERT: all-modalities-in-One BERT for multimodal sentiment analysis. Inf Fusion. 2023;92(6):37–45. doi:10.1016/j.inffus.2022.11.022.
26. Su W, Zhu X, Cao Y, Li B, Lu L, Wei F, et al. VL-BERT: pre-training of generic visual-linguistic representations. arXiv:1908.08530. 2019.
27. Tan H, Bansal M. LXMERT: learning cross-modality encoder representations from transformers. arXiv:1908.07490. 2019.
28. Zhang S, He Y, Li L, Dou Y. Multimodal sentiment analysis with BERT-ResNet50. In: Second International Conference on Algorithms, Microchips, and Network Applications (AMNA 2023); 2023 Jan 13–15; Zhengzhou, China: SPIE; 2023. 47 p. doi:10.1117/12.2679113.
29. Chandrasekaran G, Dhanasekaran S, Moorthy C, Arul Oli A. Multimodal sentiment analysis leveraging the strength of deep neural networks enhanced by the XGBoost classifier. Comput Methods Biomech Biomed Engin. 2025;28(6):777–99. doi:10.1080/10255842.2024.2313066.
30. Cai Y, Li X, Zhang Y, Li J, Zhu F, Rao L. Multimodal sentiment analysis based on multi-layer feature fusion and multi-task learning. Sci Rep. 2025;15(1):2126. doi:10.1038/s41598-025-85859-6.
31. Gan C, Tang Y, Fu X, Zhu Q, Jain DK, Garca S. Video multimodal sentiment analysis using cross-modal feature translation and dynamical propagation. Knowl Based Syst. 2024;299(01):111982. doi:10.1016/j.knosys.2024.111982.
32. Rahmani S, Hosseini S, Zall R, Kangavari MR, Kamran S, Hua W. Transfer-based adaptive tree for multimodal sentiment analysis based on user latent aspects. Knowl Based Syst. 2023;261(4):110219. doi:10.1016/j.knosys.2022.110219.
33. Wang L, Peng J, Zheng C, Zhao T, Zhu L. A cross modal hierarchical fusion multimodal sentiment analysis method based on multi-task learning. Inf Process Manag. 2024;61(3):103675. doi:10.1016/j.ipm.2024.103675.
34. Sun Z, Sarma P, Sethares W, Liang Y. Learning relationships between text, audio, and video via deep canonical correlation for multimodal language analysis. Proc AAAI Conf Artif Intell. 2020;34(5):8992–9. doi:10.1609/aaai.v34i05.6431.
35. Zadeh A, Zellers R, Pincus E, Morency LP. MOSI: multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos. arXiv:1606.06259. 2016.

36. Bagher Zadeh A, Liang PP, Poria S, Cambria E, Morency LP. Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers); 2018 Jul 15–20; Melbourne, Australia. Stroudsburg, PA, USA: ACL; 2018. p. 2236–46. doi:10.18653/v1/p18-1208.
37. Zadeh A, Liang PP, Poria S, Vij P, Cambria E, Morency LP. Multi-attention recurrent network for human communication comprehension. Proc AAAI Conf Artif Intell. 2018;32(1). doi:10.1609/aaai.v32i1.12024.
38. Pennington J, Socher R, Manning C. Glove: global vectors for word representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP); 2014 Oct 25–29; Doha, Qatar. Stroudsburg, PA, USA: ACL; 2014. p. 1532–43. doi:10.3115/v1/d14-1162.
39. Degottex G, Kane J, Drugman T, Raitio T, Scherer S. COVAREP—collaborative voice analysis repository for speech technologies. In: 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); 2014 May 4–9; Florence, Italy: IEEE; 2014. p. 960–4. doi:10.1109/icassp.2014.6853739.
40. Poria S, Cambria E, Hazarika D, Majumder N, Zadeh A, Morency LP. Context-dependent sentiment analysis in user-generated videos. In: Proceedings of the 55th Annual Meeting of the Association forComputational Linguistics (Volume 1: Long Papers); 2017 Jul 30–Aug 4; Vancouver, BC, Canada. Stroudsburg, PA, USA: ACL; 2017. p. 873–83. doi:10.18653/v1/p17-1081.
41. Tsai YH, Liang PP, Zadeh A, Morency LP, Salakhutdinov R. Learning factorized multimodal representations. arXiv:1806.06176. 2018.
42. Zadeh A, Liang PP, Mazumder N, Poria S, Cambria E, Morency LP. Memory fusion network for multi-view sequential learning. Proc AAAI Conf Artif Intell. 2018;32(1). doi:10.1609/aaai.v32i1.12021.
43. Yang D, Liu Y, Huang C, Li M, Zhao X, Wang Y, et al. Target and source modality co-reinforcement for emotion understanding from asynchronous multimodal sequences. Knowl Based Syst. 2023;265(6):110370. doi:10.1016/j.knosys.2023.110370.
44. Pham H, Liang PP, Manzini T, Morency LP, Pczos B. Found in translation: learning robust joint representations by cyclic translations between modalities. Proc AAAI Conf Artif Intell. 2019;33(1):6892–9. doi:10.1609/aaai.v33i01.33016892.
45. Wu J, Mai S, Hu H. Graph capsule aggregation for unaligned multimodal sequences. In: Proceedings of the 2021 International Conference on Multimodal Interaction; 2021 Oct 18–22; Montral, QC, Canada: ACM; 2021. p. 521–9. doi:10.1145/3462244.3479931.
46. Ma Y, Ma B. Multimodal sentiment analysis on unaligned sequences via holographic embedding. In: ICASSP 2022—2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); 2022 May 23–27; Singapore: IEEE; 2022. p. 8547–51. doi:10.1109/ICASSP43922.2022.9747646.
47. Fu Z, Liu F, Xu Q, Fu X, Qi J. LMR-CBT: learning modality-fused representations with CB-Transformer for multimodal emotion recognition from unaligned multimodal sequences. Front Comput Sci. 2023;18(4):184314. doi:10.1007/s11704-023-2444-y.
48. Rahman W, Hasan MK, Lee S, Zadeh A, Mao C, Morency LP, et al. Integrating multimodal information in large pretrained transformers. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics; 2020 Jul 6–8; Online. p. 2359–69. doi:10.18653/v1/2020.acl-main.214.
49. Hazarika D, Zimmermann R, Poria S. MISA: modality-invariant and-specific representations for multimodal sentiment analysis. In: Proceedings of the 28th ACM International Conference on Multimedia; 2020 Oct 12–16; Seattle, WA, USA: ACM; 2020. p. 1122–31. doi:10.1145/3394171.3413678.