IML-CMM - A Multimodal Sentiment Analysis Framework Integrating Intra-Modal Learning and Cross-Modal Mixup Enhancement

Zheng Zhang[1], RuiQing Yang[2], ChuanLei Zhang*
Tianjin University of Science and Technology, China
21108107@mail.tust.edu.cn[1],ruiqing.yang.study@gmail.com[2],97313114@tust.edu.cn*

Abstract

Existing multimodal sentiment analysis methods, while effective at extracting high-level unimodal features, still face challenges in the coordinated fusion of cross-modal information. These methods often struggle to fully leverage the complementary nature of different modalities. To address these limitations, this paper proposes a novel multimodal sentiment analysis framework that combines intra-modal feature learning with cross-modal mixup enhancement, termed IML-CMM. The model first utilizes KAN (Kolmogorov-Arnold Network) and Transformer to construct intra-modal feature extraction layers, which progressively embed features from text, audio, and video modalities, capturing key information at each layer. Subsequently, an Adaptive Hyper-Modality Learning (AHL) module adjusts dynamic weights between audio and video modalities, guided by multi-scale textual features, to enhance multimodal fusion. To further optimize cross-modal learning, we introduce an Audio-Visual Mixup Enhancement technique. This method mixes acoustic and visual modalities from different video sources to generate new samples, producing a consistency loss between mixed audiovisual data. The combined loss from both multimodal and mixup consistency losses is used as the overall loss, improving the model' s generalization to diverse emotional cues. Experimental results demonstrate that the proposed model achieves improvements of 1.64% to 11.14% across various metrics on the CH-SIMS v2 dataset, validating the effectiveness of its cross-modal learning capabilities.

Keywords: Multimodal Sentiment Analysis, Intra-Modal Feature Learning, Cross-Modal Fusion, Adaptive Hyper-Modality Learning, Mixup, Complementarity

1. Introduction

In today's information era, sentiment analysis, as a crucial research field, has been widely applied in areas such as social media monitoring, customer feedback analysis, and emotion-driven recommendation systems. Multimodal Sentiment Analysis (MSA), as a frontier research topic in this field, seeks to integrate information from multiple modalities such as text, audio, and visual data to enhance the accuracy and robustness of sentiment recognition. This integration not only captures emotional expressions more comprehensively but also effectively overcomes the limitations of single modalities. For example, textual information often faces ambiguity when dealing with sarcasm or implicit emotions, while audio and visual modalities can provide additional contextual information, strengthening emotional understanding through non-verbal cues such as tone and facial expressions [30, 12].

Current MSA methods typically use deep learning models like Transformers to extract high-level semantic features from individual modalities. While these models excel at capturing modality-specific features and enhancing inter-modal relationships via attention mechanisms, they face challenges in cross-modal fusion. Transformers, although effective at handling long sequences, struggle with redundant cross-modal information, leading to potential information loss [5]. Additionally, existing models often fail to fully exploit the correlations between modalities, limiting their performance in practical tasks [21].

Effective cross-modal fusion remains a major challenge. Some methods attempt direct fusion of modalities, but often fail to manage redundant or conflicting information, which degrades performance [18]. For example, the Tensor Fusion Network (TFN) introduces interactions between modalities but can create unnecessary redundancy [27]. The Adaptive Hyper-Modality Learning (AHL) module addresses this by flexibly adjusting the fusion weights of audio and video modalities, enhancing emotional cue capture [28].

However, the AHL method directly fuses audio and video modalities and heavily relies on the text modality, as text is still considered the primary source of information in many scenarios. This phenomenon is

known as the "text-predominant" problem, where the information from the audio and visual modalities is not fully utilized [20]. As a result, the single-target prediction cannot fully exploit and measure the features and contributions of the acoustic (A) and visual (V) modalities. This issue prevents the model from fully leveraging the potential of multimodal learning, thus limiting the overall performance improvement [16].

To tackle these challenges, we propose the IML-CMM framework, combining intra-modal feature learning with cross-modal mixed enhancement. The framework begins with an intra-modal feature extraction layer using KAN [24] and Transformers [8], which deeply mines features from text, audio, and video. The KAN network efficiently extracts intra-modal information, while the Transformer enhances feature embedding to capture key semantics from each modality.

Secondly, through the Adaptive Hyper-Modality Learning (AHL) module, the framework dynamically adjusts the weights of the audio and video modalities guided by multi-scale text features, enhancing the performance of multimodal fusion. This module considers both the complementarity between modalities and dynamically adjusts the contributions of the modalities based on context. However, existing AHL methods often fail to fully utilize the features and contributions of the acoustic (A) and visual (V) modalities when fusing audio and visual modalities directly.

To further optimize cross-modal learning, this paper introduces an audiovisual modality mixed enhancement (Mixup) method [12], which mixes the acoustic and visual modalities of different video sources to generate new samples. These are processed through the intra-modal feature extraction layer and a unimodal feature aggregation and classification layer constructed using Attention Pool [11] and KAN, yielding an audiovisual mixup consistency loss. This design aims to strengthen information transfer between modalities and improve the model's generalization ability to diversified emotional cues.

This study's main contributions are as follows:

1) Proposing the IML-CMM Framework: We propose the IML-CMM framework, combining intra-modal feature learning and cross-modal mixup enhancement for multimodal sentiment analysis. This framework integrates the Kolmogorov-Arnold Network (KAN) and Transformer, enhancing both intra-modal representations and cross-modal complementarity through dynamic weight adjustments and data augmentation strategies.

2) Imporving the Adaptive Hyper-Modality Learning (AHL) ModuleThe AHL module, guided by multi-scale text features, dynamically adjusts the fusion weights of audio and video modalities, addressing the "text-dominance" issue. It uses cross-modal attention and iterative feature updates to better align non-verbal cues in audio and video.

3) Developing Cross-Modal Mixed Enhancement (AV-Mixup) Technique:We combine the Mixup strategy with consistency constraints, generating new samples by mixing audio and visual data. This is paired with Mixup Consistency Loss to regularize the model, improving generalization across emotional contexts and reducing MAE by 9.43% compared to the baseline.

4) Constructing a Joint Loss Optimization Mechanism:We combine multimodal cross-entropy loss with Mixup consistency loss, balancing unimodal optimization with cross-modal constraints. This results in a 1.64% to 11.14% improvement in all metrics on the CH-SIMSv2 dataset, demonstrating the effectiveness of enhanced modal complementarity.

This research not only provides a new framework for multimodal sentiment analysis but also points the way for future studies. Future research can further explore how to optimize the interaction mechanisms between modalities to achieve more efficient emotional information fusion.

2. Related Work

In the field of Multimodal Sentiment Analysis (MSA), numerous studies have focused on integrating information from multiple modalities such as text, audio, and visual data to improve the accuracy of sentiment prediction. Early work primarily concentrated on single modalities, but with advancements in deep learning technologies, particularly in cross-modal interaction techniques, the integration of multimodal data has gradually become a research hotspot.

2.1. Multimodal Feature Extraction Techniques

Extracting meaningful features from multiple modalities is the first critical step in multimodal sentiment analysis. Traditional methods typically adopt feature concatenation or early fusion approaches, where features extracted from various modalities are combined and processed using conventional machine learning methods [15]. However, these methods often fail to capture the complex interactions between modalities, leading to suboptimal model performance.

In recent years, the Transformer architecture has gradually become the mainstream approach for fea-

ture extraction due to its ability to effectively model long-range dependencies. For example, the Multimodal Transformer (MulT) model [23] introduces a mechanism to handle both cross-modal and intra-modal dependencies, enhancing the alignment of audio and visual signals with text by learning interactions between modality pairs. Another approach, MISA [7], emphasizes extracting modality-specific features first, followed by cross-modal alignment through a shared latent space. The core idea of this model is to simultaneously retain modality-specific information and shared latent representations, thereby better fusing emotional signals from different modalities.

Additionally, adaptive techniques such as the Adaptive Hyper-Modality Learning (AHL) module [29] have been proposed, dynamically adjusting the importance of each modality based on context to address the "text-predominant" problem, where text dominates the learning process. This technology excels at balancing the contributions of audio and visual cues, especially in contexts where non-verbal data provides additional information for emotional expression.

2.2. Multimodal Fusion Strategies

Fusion strategies play a critical role in the performance of MSA models, as they determine how features from different modalities are combined to form comprehensive sentiment predictions. Early fusion strategies, such as Tensor Fusion Networks (TFN) [26], fuse the outputs of different modalities at a low level by learning tensor interactions. However, the TFN model has been criticized for introducing redundant information, particularly when handling high-dimensional feature spaces, which can lead to degraded performance.

In contrast, late fusion strategies attempt to combine the prediction results from unimodal models, often using ensemble learning methods. While this approach allows for independent processing of each modality, it overlooks potential synergies between modalities [2].

Furthermore, Mixup-based strategies, such as the Audiovisual Modality Mixup Consistent Module (AV-Mixup Consistent Module) [10], show promising results by generating new data samples through the mixing of audio and visual features from different sources. This not only improves the model's robustness but also mitigates overfitting issues on small datasets by providing a regularization technique. Our proposed IML-CMM model builds on these advanced methods, combining Transformer-based intra-modal learning with the Mixup strategy to further enhance the fusion and interaction between modalities.

3. Methodology

3.1. Overview

This section provides an overview of our proposed multimodal sentiment analysis model framework (IML-CMM), which aims to enhance sentiment analysis by combining intra-modal feature learning with cross-modal mixup enhancement. The overall architecture of the IML-CMM framework consists of the following key components:

1. Intra-modal Feature Extraction Layer: Constructed using a combination of KAN and Transformer.
2. Adaptive Hyper-Modality Learning: Utilizes multi-scale text features to guide the dynamic weight adjustment of audio and video modalities.
3. Unimodal Feature Aggregation Layer: Built using Attention Pool and KAN.
4. Cross-Modal Mixup Enhancement: Generates new samples by mixing audio and visual modalities from different video sources using the Mixup method.
5. Multimodal Loss Optimization: Combines audio-visual mixup consistency loss with the weighted sum of multimodal losses as the total loss.

3.2. Intra-modal Feature Learning

Intra-modal feature learning is a crucial step in the IML-CMM model, enabling the model to capture high-level semantic features within each modality. For the text modality, BERT is used as the feature extractor to capture contextual embeddings from the input text sequence, with the generated text feature representation denoted as $H_T \in \mathbb{R}^{n \times d_T}$.For the audio modality, Mel-frequency cepstral coefficients (MFCCs) are extracted using the Librosa library, with the audio signal's feature representation given by $H_A \in \mathbb{R}^{T_A \times F_A}$.For the video modality, facial expression and head pose features are extracted using the OpenFace toolkit, with the initial feature representation for video denoted as $H_V \in \mathbb{R}^{T_V \times F_V}$ (Kenton and Toutanova [3], 2019; McFee et al. [17], 2015; Baltrusaitis et al. [1], 2018).In the CH-SIMS dataset, the relevant dimensions for text, audio, and video are $T_V = 232$,$T_A = 925$,$F_A = 25$,$F_V = 177$ and $d_T = 768$, respectively. The extracted features from these three modalities serve as inputs to the intra-modal feature extraction layer, where Kolmogorov-Arnold Networks (KAN) act as linear connectors, mapping the nonlinear dependencies between data points within each modality. Meanwhile,

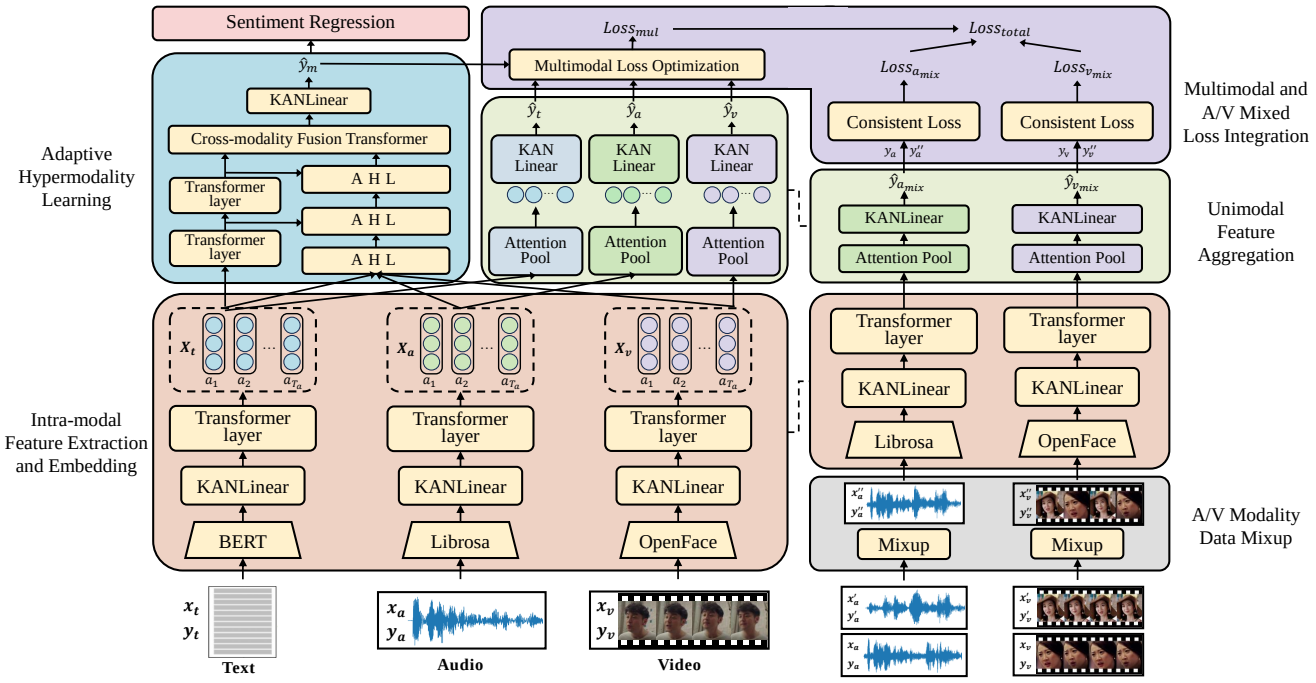

Figure 1: IML-CMM - A Multimodal Sentiment Analysis Framework Integrating Intra-Modal Learning and Cross-Modal Mixup Enhancement

the Transformer enhances the embedding of modality-specific features through a self-attention mechanism, ultimately creating an effective feature extraction process that captures both sequential dependencies and key semantic information from each modality.

3.2.1 Nonlinear Feature Mapping of KAN

Kolmogorov-Arnold Networks (KAN) are based on the Kolmogorov-Arnold theorem, utilizing combinations of univariate functions to approximate any continuous multidimensional function. The core idea of KAN networks is to perform deep nonlinear mapping of input features through nonlinear activation and basis function approximation. Specifically, KAN processes input data layer by layer through a structure of multiple KANLinear layers.

Consider an input feature matrix$X \in \mathbb{R}^{N \times d_{in}}$, where N is the number of samples and d_{in} is the dimension of the input features. In each KANLinear layer, the input data first undergoes processing through a nonlinear activation function (such as SiLU), followed by approximation using B-spline basis functions. The formula for this process is given by:

$$Y = W_{base} \cdot \sigma(X) + W_{spline} \cdot B(X), \tag{1}$$

where W_{base} is the base weight matrix, $\sigma(X)$ is the activated input features, W_{spline} is the B-spline weight matrix, and $B(X)$ represents the basis functions generated through B-spline interpolation.

B-Spline Calculation Process:

The purpose of B-spline interpolation is to provide a continuous approximation of the input space through defined grids. For each input feature x, the corresponding B-spline basis function is computed using the following recursive formula:

$$B_i(x) = \left(\frac{x - g_i}{g_{i+k} - g_i}\right) B_{i-1}(x) + \left(\frac{g_{i+k+1} - x}{g_{i+k+1} - g_{i+1}}\right) B_{i+1}(x), \tag{2}$$

where g_i represents the grid points and k is the order of the spline. By recursively calculating point by point, an approximation of the input features on the grid can be obtained.

Regularization Term:

To ensure the model's generalization ability and sparsity, KAN introduces a regularization term to constrain the smoothness of the B-spline weights. The specific regularization loss function is given by:

$$L_{reg} = \lambda_{act} \sum_{i,j} |W_{spline_{i,j}}| + \lambda_{ent} H(p), \tag{3}$$

where λ_{act} is the L1 regularization coefficient and λ_{ent} is used for entropy regularization based on weight distribution. This regularization helps the model avoid

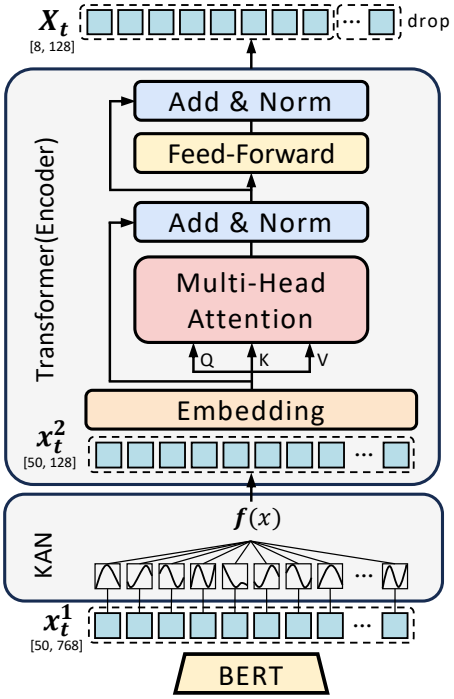

Figure 2: Intra-modal Feature Extraction Layer: Constructed using a combination of KAN and Transformer, taking the text modality features extracted by BERT as an example.

overfitting while improving the smooth approximation ability in the feature space.

In this way, the KAN network can progressively capture the nonlinear features of data when processing multimodal data through the nonlinear activation and linear combinations at each layer.

3.2.2 Transformer for Intra-Modal Embedding

After the linear transformation using KAN, the feature vectors are passed to the Transformer layer, which excels at capturing short-term and long-term dependencies within feature sequences. The standard Transformer consists of a series of multi-head self-attention layers and position-wise feed-forward networks. The self-attention mechanism calculates attention weights by measuring the similarity between elements in the sequence, making it well-suited for modeling intra-modal dependencies. For the input sequence $X = [x_1, x_2, \ldots, x_T]$, where $x_i \in \mathbb{R}^d$ represents the feature vector at time step i , the formula for calculating attention scores is:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (4)$$

where: $Q = XW_Q$, $K = XW_K$, $V = XW_V$ are the query, key, and value matrices, respectively, and the weight matrices $W_Q, W_K, W_V \in \mathbb{R}^{d \times d_k}$; d_k is the dimension of the key vectors used for scaling.

The attention output is the weighted sum of the value vectors, where the weights are determined by the similarity between the query and key vectors. The multi-head attention mechanism is defined as follows:

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \ldots, \text{head}_h)W_O \quad (5)$$

$$\text{where} \quad \text{head}_i = \text{Attention}(QW_Q^i, KW_K^i, VW_V^i) \quad (6)$$

where h represents the number of attention heads, and each head operates in parallel, allowing the model to capture different aspects of intra-modal dependencies. The output of the multi-head attention layer is passed to the feed-forward neural network:

$$\text{FFN}(x) = \text{ReLU}(xW_1 + b_1)W_2 + b_2 \quad (7)$$

where $W_1 \in \mathbb{R}^{d \times d_{\text{ff}}}$ and $W_2 \in \mathbb{R}^{d_{\text{ff}} \times d}$ are the weights of the feed-forward network, and d_{ff} is the dimension of the hidden layer.

Finally, the output of the Transformer layer is passed to subsequent layers, enabling it to capture intra-modal features at different levels.

3.3. Adaptive Hyper-Modality Learning (AHL)

The Adaptive Hyper-Modality Learning module (AHL) is the core of this framework, guiding the dynamic weight adjustment of audio and video modalities through multi-scale textual features. The AHL module consists of multiple layers of Transformer layers and AHL layers, aiming to learn language features at different scales and adaptively learn hyper-modal features from visual and audio modalities.

First, we define the textual features H_1^l as low-scale language features. Then, by introducing two Transformer layers, we learn mid-scale and high-scale language features, denoted as H_2^l and H_3^l:

$$H_i^l = \text{Transformer}(H_{i-1}^l) \quad (8)$$

where $i \in \{2, 3\}$, and H_i^l represents the language features at the ith scale.

Next, we initialize the hyper-modal features H_0^{hyper} and update H_0^{hyper} by calculating the relationship between the obtained language features and the remaining two modalities. We use a multi-head attention mechanism to compute the similarity matrix α between the language features and audio features:

$$\alpha = \text{softmax}\left(\frac{Q^l K^a}{\sqrt{d_k}}\right) \quad (9)$$

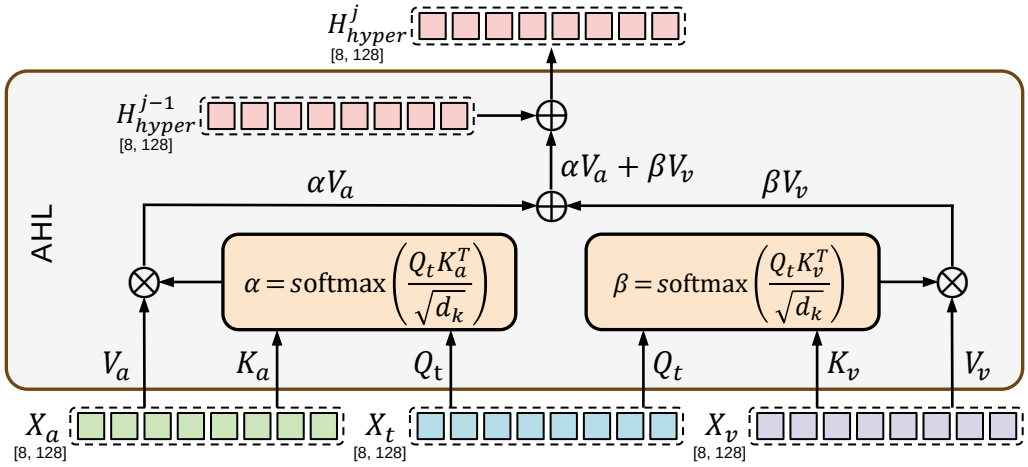

Figure 3: Adaptive Hyper-Modality Learning: Utilizes multi-scale text features to guide the dynamic weight adjustment of audio and video modalities.(First proposed by Haoyu Zhang)[28]

where Q^l and K^a are the query and key matrices, respectively, and d_k is the dimension of each attention head.

Similarly, β represents the similarity matrix between the language modality and visual modality:

$$\beta = \text{softmax}\left(\frac{Q^l K^v}{\sqrt{d_k}}\right) \tag{10}$$

Then, the hyper-modal features H_j^{hyper} can be updated using the weighted audio features and visual features:

$$H_j^{hyper} = H_{j-1}^{hyper} + \alpha V^a + \beta V^v \tag{11}$$

where $j \in \{1, 2, 3\}$, H_j^{hyper} represents the output hyper-modal features of the jth layer of AHL, and V^a and V^v are learnable weight parameters.

3.4. Unimodal Feature Aggregation(UFA)

In the Attention Pooling and KAN-based Aggregation module, we combine the attention pooling mechanism with Kolmogorov-Arnold Networks (KAN) for the aggregation and classification of unimodal features. Here, the KAN network serves as a classifier that directly classifies the pooled features without using fully connected layers. The core idea of this module is to aggregate important temporal features through attention pooling and utilize the nonlinear mapping ability of KAN to capture complex relationships between features, thereby directly predicting emotional classification.

First, let the unimodal features be represented as $H = [h_1, h_2, \ldots, h_T]$, where $h_i \in \mathbb{R}^d$ is the feature at the ith time step, T is the sequence length, and d is the feature dimension. The goal of attention pooling is to assign an attention weight α_i to each time step h_i, and generate a global feature representation by weighted summation of the feature sequence according to these weights.

The specific steps are as follows:

1. First, compute the weights for each time step feature using the attention mechanism. We calculate the attention scores for each time step using a learnable attention vector $v \in \mathbb{R}^{d_{\text{att}}}$ and a weight matrix $W_h \in \mathbb{R}^{d_{\text{att}} \times d}$:

$$e_i = v^\top \tanh(W_h h_i + b_h) \tag{12}$$

where $b_h \in \mathbb{R}^{d_{\text{att}}}$ is the bias term, and d_{att} is the hidden dimension of the attention. Then, normalize the attention scores into attention weights using the softmax function:

$$\alpha_i = \frac{\exp(e_i)}{\sum_{j=1}^{T} \exp(e_j)} \tag{13}$$

2. Use these normalized attention weights to perform a weighted summation of the feature sequence, yielding a pooled global feature vector h_{pool}:

$$h_{\text{pool}} = \sum_{i=1}^{T} \alpha_i h_i \tag{14}$$

Through attention pooling, the model can extract the most relevant features from long sequence data, generating an aggregated feature vector h_{pool} that represents global information.

The aggregated global feature h_{pool} is then input into the KAN network for classification. KAN, based on the Kolmogorov-Arnold theorem, posits that any continuous multivariate function can be decomposed into a series of simple functions. Specifically, KAN captures complex feature relationships through nonlinear transformations of the input features for emotional classification.

The computation process of KAN is as follows:

1. Perform a linear mapping on the input global feature h_{pool}:

$$y_{\text{KAN}} = W_{\text{KAN}} h_{\text{pool}} + b_{\text{KAN}} \tag{15}$$

where $W_{\text{KAN}} \in \mathbb{R}^{d' \times d}$ is the linear mapping matrix, $b_{\text{KAN}} \in \mathbb{R}^{d'}$ is the bias term, and d' is the feature dimension after mapping.

2. Then, KAN uses multilayer nonlinear transformations to capture complex relationships between features, rather than relying on simple B-spline interpolation. Through this mechanism, KAN can effectively handle complex dependencies within multimodal data.

3. Finally, KAN enhances the model's generalization ability through sparsity constraints. The regularization term is defined as:

$$L_{\text{reg}} = \lambda_{\text{act}} \sum_{i,j} \left| W_{\text{spline}_{ij}} \right| + \lambda_{\text{ent}} H(p) \tag{16}$$

where λ_{act} and λ_{ent} are the regularization coefficients, and $H(p)$ is the entropy regularization term based on the distribution of the spline weights.

Ultimately, the features processed by the KAN network, y_{KAN}, are directly used for the classification task without passing through fully connected layers. The classification results are directly given by the output of KAN:

$$\hat{y} = y_{\text{KAN}} \tag{17}$$

By combining Attention Pooling and KAN-based Aggregation, the model can extract and aggregate key features within modalities while achieving more precise emotional classification through nonlinear mapping. This design endows the classification layer with the flexibility of the attention mechanism and the powerful expressiveness of KAN in handling complex nonlinear relationships.

3.5. Cross-Modal Mixup Enhancement

In the cross-modal data augmentation section, we adopt the audiovisual mixup enhancement technique to generate new multimodal samples by mixing the acoustic and visual modalities of different instances, enhancing the model's ability to learn from unobserved multimodal contexts. This method significantly improves the model's generalization capability for multimodal information, especially when the data volume is small or the audiovisual interactions are not diverse enough.

Given a set of instances $\{(X_1, y_1), (X_2, y_2), ..., (X_n, y_n)\}$, where X_i represents the input modality (acoustic, visual, or both), and y_i is the corresponding emotion label. The basic idea of mixup is to perform linear interpolation between two randomly selected samples to generate a new mixed sample. Let' s assume we randomly select two samples X_i and X_j from the dataset, with labels y_i and y_j, respectively. The mixed sample X^{mix} and the mixed label y^{mix} are calculated as follows:

$$X^{\text{mix}} = \lambda X_i + (1 - \lambda) X_j \tag{18}$$

$$y^{\text{mix}} = \lambda y_i + (1 - \lambda) y_j \tag{19}$$

where λ is the mixing coefficient sampled from a Beta distribution, $\lambda \sim \text{Beta}(\alpha, \alpha)$, and α is a hyperparameter that controls the strength of the interpolation. Typically, λ ranges from [0, 1].

In cross-modal mixup, we mix the acoustic and visual modalities separately. Let X_{a_i} and X_{v_i} represent the acoustic and visual features of instance i, respectively. The mixing formulas for the acoustic and visual modalities are as follows:

$$X_a^{\text{mix}} = \lambda X_{a_i} + (1 - \lambda) X_{a_j} \tag{20}$$

$$X_v^{\text{mix}} = \lambda X_{v_i} + (1 - \lambda) X_{v_j} \tag{21}$$

The corresponding mixed label is:

$$y^{\text{mix}} = \lambda y_i + (1 - \lambda) y_j \tag{22}$$

By mixing the acoustic and visual modalities separately, the model can learn new audiovisual combinations, thereby enhancing its ability to handle different emotional scenarios.

To ensure that the generated mixed samples are beneficial for model training, we introduce the audiovisual mixup consistency loss, which constrains the model's predictions on the mixed samples to be consistent with the interpolated labels. Let $\hat{y}_a^{\text{mix}}$ and $\hat{y}_v^{\text{mix}}$ be the model' s predictions for the mixed acoustic and visual modalities, respectively. The audiovisual mixup consistency loss is defined as the L1 loss between the predictions and the mixed labels:

$$L_{\text{consistency}}^a = \frac{1}{N} \sum_{i=1}^{N} \left| \hat{y}_a^{\text{mix}} - y^{\text{mix}} \right| \tag{23}$$

$$L_{\text{consistency}}^v = \frac{1}{N} \sum_{i=1}^{N} \left| \hat{y}_v^{\text{mix}} - y^{\text{mix}} \right| \tag{24}$$

The final audiovisual mixup consistency loss is the weighted sum of the acoustic and visual losses:

$$L_{\text{mixup}} = \alpha L^{a}_{\text{consistency}} + \beta L^{v}_{\text{consistency}} \tag{25}$$

where α and β are weight hyperparameters used to balance the contributions of the acoustic and visual losses.

The total loss function of the model combines the multimodal loss L_{m} and the audiovisual mixup consistency loss L_{mixup}, expressed as:

$$L_{\text{total}} = L_{\text{m}} + \gamma L_{\text{mixup}} \tag{26}$$

where γ is a hyperparameter that controls the impact of the audiovisual mixup consistency loss.

The advantages of audiovisual modality mixup enhancement are reflected in several aspects: first, it effectively realizes data augmentation by generating new samples, improving the model's robustness; second, the mixup technique provides the model with more diverse audiovisual interaction combinations, enhancing its generalization ability; finally, the audiovisual mixup consistency loss ensures consistency between the predictions of mixed samples and their labels, further improving the model's performance in real-world scenarios.

3.6. Multimodal Loss Optimization

We integrate multimodal loss and audio-visual mixup consistency loss to optimize the overall performance of the multimodal sentiment analysis model. The core of this module is to optimize the model's fusion of text, audio, and video modalities through multimodal loss, while ensuring consistency between the audio and video modalities through audio-visual mixup consistency loss, thereby enhancing the model's ability to capture relevant information across modalities.

Let $\hat{y}_t$, $\hat{y}_a$, and $\hat{y}_v$ represent the classification predictions for the text, audio, and video modalities, respectively, and y_t, y_a, and y_v be the corresponding ground truth labels. The goal of the multimodal loss is to minimize the cross-entropy loss for each modality:

$$L_{\text{mul}} = L_{\text{CE}}(y_t, \hat{y}_t) + L_{\text{CE}}(y_a, \hat{y}_a) + L_{\text{CE}}(y_v, \hat{y}_v) \tag{27}$$

where the cross-entropy loss L_{CE} is defined as:

$$L_{\text{CE}}(y, \hat{y}) = -\sum_{c=1}^{C} y_c \log \hat{y}_c \tag{28}$$

Here, C is the number of classes, y_c is the label value for class c in the ground truth, and $\hat{y}_c$ is the predicted probability for class c by the model.

By minimizing the losses for the text, audio, and video modalities individually, the model can better integrate information from each modality in the multimodal sentiment analysis task.

To further enhance the correlation between the audio and video modalities, we adopt the Audio-Visual Mixup Consistency Loss. The purpose of this loss is to ensure that the new samples generated through mixing maintain consistency between the audio and video modalities.

1. Mixup Data Generation: First, we randomly mix the original audio and video data (x_a, x_v) using weighted combinations to generate new audio and video data (x'_a, x'_v). The specific generation formula is as follows:

$$x'_a = \lambda x_a + (1-\lambda) x''_a \tag{29}$$

$$x'_v = \lambda x_v + (1-\lambda) x''_v \tag{30}$$

where x''_a, x''_v are the audio and video data from other samples, and $\lambda \in [0, 1]$ is a mixing coefficient sampled from a Beta distribution.

2. Mixup Consistency Loss: For the mixed audio and video data, we calculate the predicted results $\hat{y}'_a$ and $\hat{y}'_v$ using the model. We aim to maintain consistency between the modalities for the new data generated through mixing, thus introducing a consistency loss to constrain the predictions for the mixed data:

$$L_{\text{mix}} = \lambda L_{\text{CE}}(y_a, \hat{y}'_a) + (1-\lambda) L_{\text{CE}}(y_v, \hat{y}'_v) \tag{31}$$

Here, L_{CE} remains the cross-entropy loss, and λ controls the weight balance between the audio and video modalities.

To optimize both multimodal fusion and audiovisual mixup consistency simultaneously, the final total loss function L_{total} is defined as:

$$L_{\text{total}} = L_{\text{mul}} + L_{\text{amix}} + L_{\text{vmix}} \tag{32}$$

where L_{mul} is the multimodal loss, and L_{amix} and L_{vmix} are the mixup consistency losses for audio and video, respectively. Through this integrated loss function, the model not only performs well across all modalities but also improves overall performance by enforcing consistency constraints between the audio and visual modalities.

By integrating the above losses, the model can fully utilize the features of each modality while enhancing the interrelations between audio and video modalities, thus improving the accuracy of multimodal sentiment classification.

Item	Type	Total	NEG	WNEG	NEU	WPOS	POS
#Train	Supervised	2722	921	433	232	318	818
#Valid	Supervised	647	224	110	62	83	168
#Test	Supervised	1034	291	211	93	183	256
#Unsupervised	Unsupervised	10161	-	-	-	-	-

Table 1: Data splits in the CH-SIMS v2.0 dataset. NEG: Negative, WNEG: Weak Negative, NEU: Neutral, WPOS: Weak Positive, POS: Positive.[28]

4. Experiments

4.1. Datasets

This experiment uses the CH-SIMSv2 Chinese multimodal sentiment analysis dataset, which includes text, audio, and video data for emotion classification. Each sample consists of Chinese dialogues with emotion labels, covering a range from positive to negative, and includes semantic text, speech, and facial video features. The goal is to improve emotion recognition by fusing multimodal information.

In addition to the labeled data, the dataset also contains an unsupervised set of 10,161 raw video segments (see Table 2), which can be used for tasks like representation learning, pretraining, and self-supervised approaches. This large-scale, unlabeled data provides flexibility for training models and improving emotion recognition system robustness.

4.2. Evaluation Metrics

To evaluate model performance, we use several common metrics: Acc2 measures binary classification accuracy, while F1-score balances precision and recall. Acc2-Weak evaluates performance with weak labels, and Corr (Pearson Correlation) reflects the model's ability to capture emotional fluctuations. R-square indicates the model' s goodness of fit, with larger values showing better alignment with emotional features. MAE measures the error magnitude between predicted and true values, with smaller values indicating better performance.

4.3. Baselines

To validate the effectiveness of our model, we selected several existing multimodal sentiment analysis models as baselines for comparison, including both single-target and multi-target models:

LF_DNN[14], TFN (Tensor Fusion Network)[26], LMF (Low-rank Multimodal Fusion)[14], MFN (Memory Fusion Network)[4], Graphn_MFN, MulT (Multimodal Transformer)[22], Bert_MAG (BERT with Modality Attention Gate)[19], MISA (Modality Invariant and Specific Representations)[9], MMIM (Multimodal Information Bottleneck)[6], Self_MM[25], ALMT (Adaptive Learning Multimodal Transformer).

The model proposed in this paper is implemented in Pytorch 2.0.0, and the epoch is trained around 100 times on NVIDIA RTX 3090 GPU with an initial learning rate of 1e-5.

4.4. Performance Comparison

Through a detailed analysis of the comparative performance of the proposed IML-CMM model against various established multimodal sentiment analysis models, the results indicate that IML-CMM outperforms AV-MC and other prominent models across multiple evaluation metrics.

In terms of Acc2 and F1-score, IML-CMM achieved the highest scores of 83.85% and 83.95%, respectively, surpassing AV-MC by 1.64% and 1.70%. This performance demonstrates that IML-CMM is more effective in aligning sentiment predictions with true values, showing significant improvement over other models such as MISA, MMIM, and Bert_MAG.

For the Acc2-weak metric, IML-CMM achieved a score of 76.89%, which is a notable increase of 3.15% compared to AV-MC. This result indicates that IML-CMM is more robust when dealing with weaker signals or noisy data, confirming its enhanced generalization capability in challenging scenarios.

In terms of correlation (Corr) and R-squared, IML-CMM also excelled, achieving a Corr of 76.10% (an increase of 4.00%) and an R-squared of 56.29% (an increase of 11.14%). These results reflect the model's superior ability to capture the linear relationship between predicted and actual values, further improving the accuracy of sentiment predictions.

Finally, IML-CMM achieved the lowest mean absolute error (MAE), reducing it by 9.43% compared to AV-MC, with a value of 0.269. This highlights the

Models	Acc2 (↑)	F1-score (↑)	Acc2-weak (↑)	Corr (↑)	R-squre (↑)	MAE (↓)
LF_DNN	73.95	73.84	69.13	52.19	20.84	0.381
TFN	76.51	76.31	66.27	66.65	35.9	0.323
LMF	77.05	77.02	69.34	63.75	40.64	0.343
MFN	75.27	75.24	66.46	60.6	32.26	0.355
Graphn_MFN	73.98	73.62	69.82	49.71	13.78	0.396
MulT	79.5	79.59	69.61	70.32	47.15	0.317
Bert_MAG	79.79	79.78	71.87	69.09	43.08	0.334
MISA	80.53	80.63	70.5	72.49	50.59	0.314
MMIM	80.95	80.97	72.28	70.65	43.81	0.316
Self_MM	79.01	78.89	71.87	64.03	29.36	0.335
ALMT	81.19	81.57	/	61.9	/	0.404
MLF_DNN*	78.4	78.44	71.59	65.8	39.34	0.326
MTFN*	80.26	80.33	71.07	70.54	46.07	0.318
MLMF*	79.92	79.72	69.88	71.37	47.53	0.302
AV-MC*	82.5	82.55	74.54	73.17	50.65	0.297
IML-CMM(Ours)	83.85(1.64%)	83.95(1.70%)	76.89(3.15%)	76.1(4.00%)	56.29(11.14%)	0.269(9.43%)

Table 2: Model performances for Traditional Multimodal Sentiment Analysis model on the CH-SIMS v2.0 dataset. Models with * are trained on multitasking. ALMT benchmark results are drawn from Zhang et al. (2023) [28], while the other model performances are based on the dataset introduced by Liu et al. (2022) [13].

model's improved accuracy in predicting continuous-value sentiment labels.

In summary, the IML-CMM framework consistently outperformed baseline models, demonstrating better accuracy, generalization capability, and error minimization, validating its effectiveness in multimodal sentiment recognition tasks.

On the unsupervised multi-scene video dataset with 10161 segments, our framework demonstrates strong performance. The results show Acc2 of 83.75%, Acc2-weak of 76.19%, Corr of 75.86%, and a low MAE of 26.98. These metrics highlight the framework's effectiveness in both sentiment polarity classification and fine-grained intensity prediction. The robust performance on this challenging dataset further validates the generalizability of our framework for multimodal sentiment analysis.

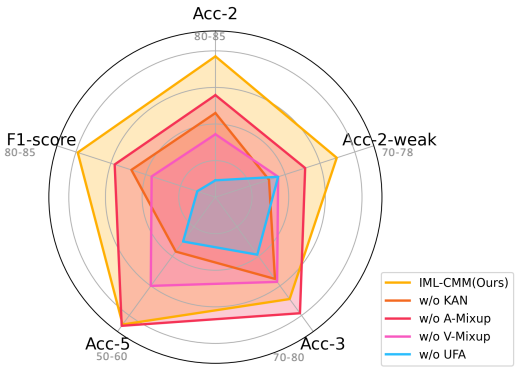

Figure 4: Performance Metrics Comparison of Different Modality Combinations in the IML-CMM Model

5. Ablation Study and Analysis

5.1. Impact of Different Components

To verify the contribution of each component of the IML-CMM model to overall performance, we conducted detailed ablation experiments to observe changes in model performance after removing each component, revealing the importance of each component. The following discussion focuses on the impact of KAN, UFA (Unimodal Feature Aggregation layer), Mixup-A, Mixup-V, and AHL (Adaptive Hyper-Modality Learning module) on model performance,As shown in Table 4.

Replacing the Knowledge Attention Network (KAN) with a standard Multi-Layer Perceptron (MLP) caused notable performance drops, especially in accuracy (Acc-2), F1-score, and correlation (Corr). KAN' s complex attention mechanism captures subtle cross-modal semantic features, which MLP fails to do, leading to a loss in the model's ability to integrate crucial knowledge-based information. This reinforces the value of advanced attention mechanisms in multimodal sentiment analysis.

The Unimodal Feature Aggregation (UFA) layer,

Models	Acc2 (↑)	F1-score (↑)	Acc2-weak (↑)	Corr (↑)	R-squre (↑)	MAE (↓)
IML-CMM(Ours)	83.75	83.81	76.19	75.86	53.75	0.27

Table 3: Performance on the multi-scene video dataset with 10161 segments

Method	Acc-2	Acc-2-weak	Acc-3	Acc-5	F1-score	MAE	Corr	R-squre	Loss
IML-CMM(Ours)	83.85	75.57	76.89	58.61	83.95	0.269	76.1	56.29	0.713
w/o KAN (MLP instead)	82.3	72.46	75.53	53.7	82.41	0.286	75.22	55.86	0.743
w/o UFA	80.46	72.88	73.89	53	80.52	0.305	71.07	48.5	0.298
w/o Mixup-A	82.79	74.12	77.85	58.7	82.89	0.274	75.4	53.53	0.733
w/o Mixup-V	81.72	72.88	75.73	56	81.83	0.291	73.2	51.33	0.816

Table 4: Impact of Component Removal on Performance Metrics in IML-CMM

which aggregates key information across modalities, also showed significant impact. Removing UFA led to a drop in both Acc-2 and F1-score, with MAE increasing, demonstrating the importance of unimodal feature extraction in maintaining cross-modal consistency. The UFA layer ensures accurate unimodal predictions, laying a strong foundation for successful multimodal fusion.

The Mixup enhancements for audio (Mixup-A) and video (Mixup-V) modalities further highlighted the importance of modality-specific data augmentation. Removing Mixup-A caused a slight decline in performance, but removing Mixup-V resulted in a more pronounced drop, particularly in Acc-2 and F1-score. This suggests that mixing video information plays a more vital role in improving model performance than audio.

Finally, removing the Adaptive Hyper-Modality Learning module (AHL) led to a dramatic decline across all metrics, with Acc-2 and F1-score falling sharply, and R-square even turning negative. AHL dynamically adjusts modality weights to optimize the fusion of audio and video information, significantly enhancing the model' s ability to capture emotional cues across modalities. Without AHL, the model struggles with cross-modal information fusion, highlighting its critical role in multimodal sentiment analysis.

In conclusion, the ablation experiments demonstrate that KAN, UFA, Mixup strategies, and AHL are crucial for the IML-CMM model' s high performance. Particularly, AHL ensures effective fusion and dynamic adjustment between modalities, enabling the model to better capture complex emotional cues across different data types.

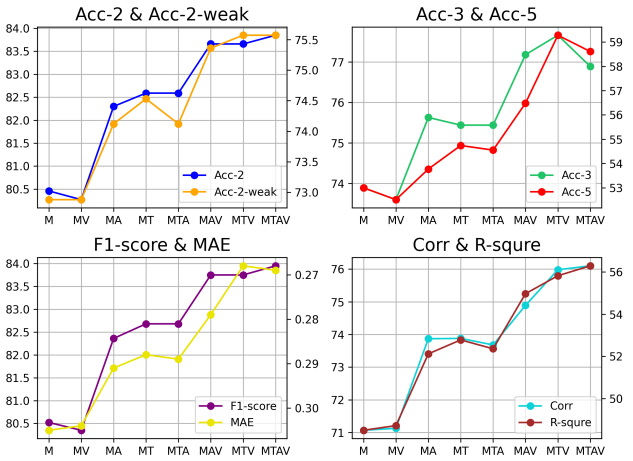

Figure 5: Performance Metrics Comparison of Different Modality Combinations in the IML-CMM Model

5.2. Effects of Different Modalities

In multimodal sentiment analysis (MSA), the contribution of each modality is crucial to the overall performance of the model. The experimental results indicate that the combination of different modalities significantly influences key metrics such as accuracy, F1-score, MAE, and correlation.

First, the text modality (T) often plays a dominant role in sentiment analysis because it directly conveys semantic information. The experiments show that when using only the text modality (MT), the model achieves an Acc-2 of 82.59% and an F1-score of 82.68%, demonstrating strong performance. This underscores the central role of text in capturing emotional cues. However, relying solely on text may lead to ambiguities in cases of sarcasm or metaphor, where the emotional intent is harder to discern from text alone.

Modal	Acc-2	Acc-2-weak	Acc-3	Acc-5	F1-score	MAE	Corr	R-squre	Loss
M	80.46	72.88	73.89	53	80.52	0.305	71.07	48.5	0.298
MV	80.27	72.88	73.6	52.51	80.35	0.304	71.13	48.72	0.258
MT	82.59	74.53	75.44	54.74	82.68	0.288	73.88	52.78	0.332
MA	82.3	74.12	75.63	53.77	82.36	0.291	73.87	52.12	0.717
MTA	82.59	74.12	75.44	54.55	82.68	0.289	73.68	52.36	0.651
MAV	83.66	75.36	77.18	56.48	83.75	0.268	74.89	54.97	1.059
MTV	83.66	75.57	77.66	59.28	83.75	0.268	75.98	55.82	0.483
MTAV	83.85	75.57	76.89	58.61	83.95	0.269	76.1	56.29	0.713

Table 5: Performance Metrics of Different Modality Combinations in IML-CMM

The audio modality (A) provides complementary non-verbal cues, such as tone and speech rate, which are essential for emotion recognition. Although the Acc-2 of the audio modality alone (MA) is 82.30%, slightly lower than that of text, it still significantly enhances the model, particularly in speech-driven contexts, where audio captures nuances that text might miss.

The video modality (V), when used independently (MV), shows relatively weaker performance with an Acc-2 of 80.27%. This suggests that relying solely on visual features for emotion recognition is somewhat limited. However, video can provide critical emotional information through facial expressions and body language, which plays a valuable role in multimodal combinations.

When text, audio, and video modalities are combined (MTAV), the model' s performance significantly improves, achieving an Acc-2 of 83.85% and an F1-score of 83.95%. This demonstrates that multimodal fusion can effectively address the shortcomings of individual modalities, leading to a more comprehensive understanding of emotional cues. Furthermore, the use of audiovisual mixing enhancement (Mixup) generates new audiovisual samples, improving the model' s generalization and adaptability to diverse emotional expressions. Overall, the joint use of multimodal data significantly boosts emotion recognition effectiveness, validating the strengths and advancements of the IML-CMM model in cross-modal learning.

5.3. Impact of Modality Weights

In Section 5.2, we discussed the roles of different modalities in emotional analysis, highlighting the importance of audio, video, and other modalities in multimodal fusion. Next, we analyze how modal weight allocation affects overall performance based on 100 experimental sets with varying MTAV (Multimodal Weights: M, T, A, V), randomly selected from the range [0.2, 0.4, 0.6, 0.8, 1].

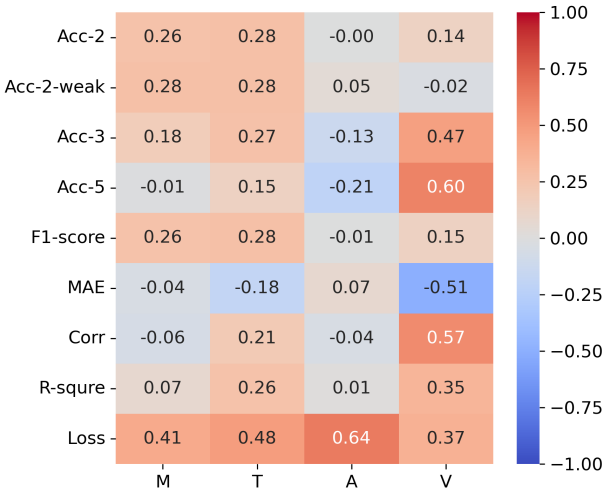

Figure 6: Correlation Heatmap of Modal Weights Impact on Model Performance

The heatmap reveals that modal weights significantly impact correlation levels across evaluation metrics. Acc-2 shows a positive correlation with the weights of M (fusion modality) and T (text modality) (0.26 and 0.28, respectively), suggesting that increasing these weights enhances binary classification accuracy. The weight of A (audio modality) has a smaller effect on Acc-2.

For Acc-5, the weight of the video modality V has a strong positive correlation (0.60), indicating its importance in fine-grained emotion classification tasks. The weight of M also positively impacts Acc-5 and F1-score (0.26 for both), highlighting the effectiveness of the AHL-based dynamic fusion in capturing modality synergies.

Regarding MAE, there is a significant negative correlation with V (-0.51), meaning increasing video weight reduces prediction error. Text modality T also shows a negative correlation with MAE (-0.18), reinforcing its role in precise emotional feature extraction. Changes in the weight of M have minimal impact on MAE (-0.04), indicating that while the fusion modality aggregates information, its direct effect on error is small.

In conclusion, these experiments show that the dynamic adjustment of fusion modality M enhances performance, particularly in classification accuracy and F1-score. Video modality weight helps fine-grained emotion classification, while text plays a key role in reducing errors. However, increasing the audio modality weight excessively may lead to higher training loss.

6. Conclusion

The IML-CMM framework proposed in this paper combines intra-modal feature learning with cross-modal mixing enhancement, addressing the limitations of existing methods in the fusion of information between modalities. By jointly utilizing KAN and Transformer, IML-CMM can effectively capture high-level features of text, audio, and video modalities, while the nonlinear mapping of KAN effectively resolves the complexity of information extraction within modalities. The Adaptive Hyper-Modality Learning (AHL) module leverages multi-scale text features to guide the dynamic weight adjustments of audio and video modalities, enhancing the performance of multimodal fusion. The audiovisual modality mixing enhancement (Mixup) strategy generates new samples by mixing features from different audio and video samples, enabling the model to better capture emotional cues within the audio and video modalities. Meanwhile, IML-CMM employs an optimization scheme that combines multimodal loss with audiovisual mixing consistency loss. The multimodal loss optimizes the independent features of each modality, while the audiovisual mixing consistency loss constrains the predictions of mixed samples to ensure the consistency and complementarity of information across different modalities. Experimental results show that IML-CMM achieves performance improvements of 1.64% to 11.14% on the CH-SIMSv2 dataset compared to existing models, particularly excelling in key metrics such as accuracy and correlation. Results from the ablation studies further validate the crucial contributions of the Mixup enhancement strategy, UFA module, and AHL module to the model's performance. This not only confirms the effectiveness of the IML-CMM framework in intra-modal learning and cross-modal mixing enhancement but also highlights its applicability in complex emotional scenarios. Future work can continue to explore optimization strategies and loss designs for multimodal interactions to enhance the model's applicability in even more complex emotional contexts.

References

[1] T. Baltrušaitis, C. Ahuja, and L.-P. Morency. Multimodal machine learning: A survey and taxonomy. volume 41, pages 423–443. IEEE, 2018. 3

[2] R. Chen, W. Zhou, H. Hu, Z. Fei, M. Fei, and H. Zhou. Disentangled variational auto-encoder for multimodal fusion performance analysis in multimodal sentiment analysis. volume 301, page 112372. Elsevier, 2024. 3

[3] J. Devlin. Bert: Pre-training of deep bidirectional transformers for language understanding. 2018. 3

[4] P. Fu, J. Wang, X. Zhang, L. Zhang, and R. X. Gao. Dynamic routing-based multimodal neural network for multi-sensory fault diagnosis of induction motor. volume 55, pages 264–272. Elsevier, 2020. 9

[5] X. Gu, P. P. Liang, and L.-P. Morency. Multimodal sentiment analysis using deep learning. In Proceedings of the AAAI Conference on Artificial Intelligence, 2020. 1

[6] W. Han, H. Chen, and S. Poria. Improving multimodal fusion with hierarchical mutual information maximization for multimodal sentiment analysis. arXiv preprint arXiv:2109.00412, 2021. 9

[7] D. Hazarika, R. Zimmermann, and S. Poria. Misa: Modality-invariant and -specific representations for multimodal sentiment analysis. In Proceedings of the 28th ACM International Conference on Multimedia, MM '20, page 1122–1131, New York, NY, USA, 2020. Association for Computing Machinery. 3

[8] D. Hazarika, R. Zimmermann, and S. Poria. Misa: Modality-invariant and specific representations for multimodal sentiment analysis. In ACM International Conference on Multimedia, 2020. 2

[9] D. Hazarika, R. Zimmermann, and S. Poria. Misa: Modality-invariant and-specific representations for multimodal sentiment analysis. In Proceedings of the 28th ACM international conference on multimedia, pages 1122–1131, 2020. 9

[10] X. Jin, H. Zhu, S. Li, Z. Wang, Z. Liu, C. Yu, H. Qin, and S. Z. Li. A survey on mixup augmentations and beyond. 2024. 3

[11] K. Kim and S. Park. Aobert: All-modalities-in-one bert for multimodal sentiment analysis. Information Fusion, 2023. 2

[12] Y. Liu, Z. Yuan, H. Mao, Z. Liang, W. Yang, Y. Qiu, T. Cheng, X. Li, H. Xu, and K. Gao. Make acoustic and visual cues matter: Ch-sims v2.0 dataset and av-mixup consistent module. In Proceedings of the 2022 International Conference on Multimodal Interaction, ICMI '22, page 247–258, New York, NY, USA, 2022. Association for Computing Machinery. 1, 2

[13] Y. Liu, Z. Yuan, H. Mao, Z. Liang, W. Yang, Y. Qiu, T. Cheng, X. Li, H. Xu, and K. Gao. Make acoustic and visual cues matter: Ch-sims v2.0 dataset and av-mixup consistent module. 2022. 10

[14] Z. Liu, Y. Shen, V. B. Lakshminarasimhan, P. P. Liang, A. Zadeh, and L.-P. Morency. Efficient low-rank multimodal fusion with modality-specific factors. 2018. 9

[15] F. Ma, Y. Zhang, and X. Sun. Multimodal sentiment analysis with preferential fusion and distance-aware contrastive learning. In 2023 IEEE International Conference on Multimedia and Expo (ICME), pages 1367–1372. IEEE, 2023. 2

[16] S. Mai, H. Hu, and S. Xing. Hierarchical feature fusion for multimodal sentiment analysis. In Proceedings of ACL Conference, 2019. 2

[17] B. McFee, C. Raffel, D. Liang, D. P. Ellis, M. McVicar, E. Battenberg, and O. Nieto. librosa: Audio and music signal analysis in python. In SciPy, pages 18–24, 2015. 3

[18] H. Pham, P. P. Liang, and L.-P. Morency. Found in translation: Learning robust joint representations by cyclic translations between modalities. In AAAI Conference on Artificial Intelligence, 2019. 1

[19] W. Rahman, M. K. Hasan, S. Lee, A. Zadeh, C. Mao, L.-P. Morency, and E. Hoque. Integrating multimodal information in large pretrained transformers. In Proceedings of the conference. Association for Computational Linguistics. Meeting, volume 2020, page 2359. NIH Public Access, 2020. 9

[20] Z. Sun, P. Sarma, W. Sethares, and Y. Liang. Learning relationships between text, audio, and video via deep canonical correlation for multimodal language analysis. In AAAI Conference on Artificial Intelligence, 2020. 2

[21] Y.-H. Tsai, P. P. Liang, A. Zadeh, L.-P. Morency, and R. Salakhutdinov. Multimodal transformer for unaligned multimodal language sequences. In Proceedings of the ACL Conference, 2019. 1

[22] Y.-H. H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L.-P. Morency, and R. Salakhutdinov. Multimodal transformer for unaligned multimodal language sequences. In Proceedings of the conference. Association for computational linguistics. Meeting, volume 2019, page 6558. NIH Public Access, 2019. 9

[23] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. In Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS'17, page 6000–6010, Red Hook, NY, USA, 2017. Curran Associates Inc. 3

[24] Y. Wang, Y. Shen, Z. Liu, P. P. Liang, A. Zadeh, and L.-P. Morency. Tedt: Transformer-based encoding-decoding translation network for multimodal sentiment analysis. Cognitive Computation, 2021. 2

[25] W. Yu, H. Xu, Z. Yuan, and J. Wu. Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis. In Proceedings of the AAAI conference on artificial intelligence, volume 35, pages 10790–10797, 2021. 9

[26] A. Zadeh, M. Chen, S. Poria, E. Cambria, and L.-P. Morency. Tensor fusion network for multimodal sentiment analysis. 2017. 3, 9

[27] A. Zadeh, P. P. Liang, S. Poria, E. Cambria, and L.-P. Morency. Tensor fusion network for multimodal sentiment analysis. In Proceedings of EMNLP, 2017. 1

[28] H. Zhang, Y. Wang, G. Yin, K. Liu, Y. Liu, and T. Yu. Learning language-guided adaptive hyper-modality representation for multimodal sentiment analysis. In H. Bouamor, J. Pino, and K. Bali, editors, Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pages 756–767, Singapore, Dec. 2023. Association for Computational Linguistics. 1, 6, 9, 10

[29] H. Zhang, Y. Wang, G. Yin, K. Liu, Y. Liu, and T. Yu. Learning language-guided adaptive hyper-modality representation for multimodal sentiment analysis. 2023. 3

[30] K. Zhang, Y. Geng, J. Zhao, J. Liu, and W. Li. Sentiment analysis of social media via multimodal feature fusion. Symmetry, 12(12):2010, 2020. 1