Journal of Intelligent Information Systems
https://doi.org/10.1007/s10844-025-01006-7

RESEARCH

# Frequency-aware experts with multi-stage fusion for multimodal sentiment analysis

**Xiaofei Zhu**[1] · **Yaochen Li**[1]

Received: 21 August 2025 / Revised: 5 November 2025 / Accepted: 7 November 2025


**Abstract**
Multimodal Sentiment Analysis (MSA) aims to infer human affective states by integrating information from diverse modalities such as text, audio, and vision. Despite recent advances in representation learning and fusion strategies, existing methods often overlook the inherent frequency characteristics within each modality–particularly in audio and visual signals–where task-relevant information may reside in distinct spectral bands. To address this limitation, we propose a novel Frequency-Aware Experts and Multi-Stage Fusion (FEMF) framework. Specifically, we introduce a frequency-aware expert module that decomposes modality-private features into high- and low-frequency components via Discrete Fourier Transform (DFT), and processes them through dedicated expert networks before adaptive fusion. Additionally, we design a multi-stage integration pipeline that incorporates shared-private disentanglement, multi-query modality interaction, and confidence-aware fusion with hierarchical prediction, enabling flexible and robust representation learning across modalities. Extensive experiments on CMU-MOSI and CMU-MOSEI benchmarks demonstrate that our approach achieves superior performance, validating the effectiveness and resilience of the proposed frequency-aware modeling paradigm. Codes are realised at https://github.com/L11yc/FEMF

**Keywords** Multimodal Sentiment Analysis · Frequency-Aware Modeling · Mixture of Experts · Multi-Stage Fusion

## 1 Introduction

With the rapid proliferation of social media platforms, users increasingly express emotions through diverse modalities such as text, audio, and video. This has led to the rise of Multimodal Sentiment Analysis (MSA), which aims to infer human affective states by modeling

✉ Xiaofei Zhu
zxf@cqut.edu.cn

Yaochen Li
lyc@stu.cqut.edu.cn

[1] College of Computer Science and Engineering, Chongqing University of Technology, Hongguang Avenue, 400054 Chongqing, China

Published online: 18 November 2025

heterogeneous multimodal data sources (Wang et al., 2025a). MSA has demonstrated wide applicability in domains such as healthcare, social media monitoring, and human–computer interaction (Yang et al., 2023a). Compared with unimodal approaches, MSA provides superior robustness and richer semantic understanding by leveraging complementary cross-modal cues.

Early studies primarily relied on textual data, while the increasing availability of user-generated videos has shifted attention toward fully exploiting multimodal representations. Recent works have investigated representation learning to improve the expressiveness and disentanglement of modality-shared and modality-specific features (Hazarika et al., 2020), employing strategies such as shared–private factorization, contrastive learning (Yang et al., 2023a), and multitask learning (Yu et al., 2021). In parallel, fusion strategies have evolved from early concatenation to more expressive mechanisms, including tensor-based fusion (Zadeh et al., 2017a), convolutional and recurrent architectures (Sun et al., 2020; Huang et al., 2020), and cross-modal attention networks (Li et al., 2025a). However, most existing methods overlook intrinsic frequency characteristics within modalities: while audio signals exhibit rich temporal variations across frequency bands, visual and textual features often contain low-frequency semantic trends interspersed with sparse high-frequency fluctuations. Directly modeling these heterogeneous frequency patterns in the original feature space may dilute task-relevant information and obscure subtle but crucial modality dynamics (Ai et al., 2025).

To address these challenges, we propose FEMF, a Frequency-Aware Experts with Multi-Stage Fusion framework for multimodal sentiment analysis. Motivated by the observation that different frequency bands encode distinct emotional cues—high-frequency components capturing abrupt emotional transitions and low-frequency ones reflecting stable affective trends (Cheng et al., 2025)—FEMF decomposes modality-specific features via Discrete Fourier Transform (DFT) into high- and low-frequency components, which are processed by specialized experts and adaptively fused to form enhanced representations. This frequency-aware design preserves both global emotional consistency and transient variations, mitigating the semantic blurring caused by ignoring frequency cues. Moreover, FEMF integrates shared–private disentanglement guided by cross-modal reconstruction to enhance multimodal representation learning and promote robust fusion.

Extensive experiments on CMU-MOSI and CMU-MOSEI benchmarks demonstrate the effectiveness and robustness of our approach. The results validate the utility of frequency-aware expert modeling and highlight its potential for interpretable and resilient multimodal sentiment analysis.

The main contributions of this work are as follows:

1. We present a comprehensive Frequency-Aware Experts with Multi-Stage Fusion (FEMF) framework comprising four key stages: shared–private disentanglement, frequency-aware expert modeling, modality interaction, and confidence-aware fusion with hierarchical prediction. This structured pipeline enables expressive, interpretable, and robust multimodal representation learning.
2. We introduce a novel frequency-aware expert module that decomposes modality-private features via DFT into high- and low-frequency components, which are processed by separate experts and adaptively fused to form enriched representations.

3. Extensive experiments on CMU-MOSI and CMU-MOSEI show that FEMF outperforms strong baselines under both full-modality and missing-modality settings, validating the generalizability of frequency-aware modeling for multimodal affective understanding.

## 2 Related work

### 2.1 Disentangled multimodal representation learning

While disentangled multimodal representation learning has shown promise in separating shared and modality-specific information (Zhang et al., 2024a), existing approaches still face key challenges. Early symmetric methods (Hazarika et al., 2020) treat all modalities equally, overlooking modality dominance and often diluting strong cues such as language. Distillation-based frameworks like DMD (Li et al., 2023) and CorrKD (Li et al., 2024a) align global distributions but neglect fine-grained or hierarchical semantic consistency. Recent models, including What2Comm (Yang et al., 2023c) and Tailor (Zhang et al., 2022), further ignore dominant-modality-guided disentanglement, limiting performance under asymmetric information.

In summary, existing methods, though technically advanced, lack dominant modality awareness, effective filtering of redundant features, and semantic consistency across hierarchies–highlighting the need for a disentanglement strategy that enhances key modalities while ensuring layered semantic alignment.

### 2.2 Mixture of experts

The Mixture of Experts (MoE) architecture has recently emerged as an effective paradigm for enhancing model capacity and representation learning in natural language processing (Lepikhin et al., 2021) and computer vision (Mustafa et al., 2022). A typical MoE consists of multiple expert subnetworks and a routing mechanism that dynamically selects relevant experts based on the input. Compared with static fusion models, MoE provides greater adaptability by allowing each expert to specialize in distinct subspaces while the router learns task-relevant expert selection (Mustafa et al., 2022). For instance, the Switch Transformer (Lepikhin et al., 2021) improves computational efficiency by routing tokens to a small subset of experts. Recent multimodal extensions design modality-specific experts and routing functions to determine their relative contributions. MoMKE (Xu et al., 2024), for example, employs a soft routing module to blend modality-specific and cross-modal representations, enabling the model to capture both unique and shared semantics across modalities.

Overall, the MoE-based architectures offer a promising alternative to conventional multimodal fusion strategies. Their ability to perform dynamic expert selection, handle diverse input conditions, and exploit specialization makes them particularly suitable for complex multimodal tasks where static fusion mechanisms fall short.

### 2.3 Fourier domain learning

Frequency-domain representations have recently gained attention as an effective means of capturing periodic and hierarchical patterns in multimodal signals. Unlike time-domain processing, frequency-based modeling decomposes data into components with distinct spectral characteristics, enabling selective emphasis on stable or transient variations (Cheung et al., 2020). This paradigm has proven valuable in computer vision for enhancing texture perception and suppressing feature-map noise (Zhang et al., 2025, 2019), in natural language processing for modeling global semantics beyond sequential context (Ong & Khong, 2025), and in time-series learning for improving long-term dependency modeling and alleviating Transformer over-smoothing (Shi et al., 2024).

Recently, frequency-domain techniques have been extended to affective computing, where emotional signals inherently exhibit rich temporal–spectral structures (Ai et al., 2025). For example, high-frequency speech cues often correspond to transient emotions like anger or surprise, whereas low-frequency contours capture stable affective states. Similarly, facial expressions contain both rapid micro-expressions and gradual transitions across distinct spectral bands (Cheng et al., 2025). Motivated by these insights, our work integrates frequency-aware modeling into multimodal sentiment analysis to disentangle and exploit high- and low-frequency affective cues, thereby enhancing interpretability and robustness in emotion understanding.

### 2.4 Multimodal fusion methods

Multimodal fusion aims to construct high-quality joint representations through effective integration across modalities. Despite notable progress, existing methods still struggle to model the complex and heterogeneous nature of cross-modal interactions. Most approaches emphasize atomic-level correlations while overlooking composition-level and hierarchical semantic structures (Liu et al., 2024; Wu et al., 2025; Li et al., 2025b, 2024b). Attention-based and graph-structured models capture local dependencies (Zhang et al., 2024b) but lack mechanisms for modeling hierarchical relationships across modalities. Moreover, modality-specific encoders often introduce redundant or irrelevant content–especially in visual streams–leading to noisy fused representations (Li et al., 2022). Although high-capacity tensor or polynomial fusion frameworks offer richer expressiveness, they suffer from excessive computational cost and poor scalability (Zadeh et al., 2017b; Hou et al., 2019). Even recent methods employing mutual information maximization, top-down feedback (Paraskevopoulos et al., 2022) fail to capture bidirectional semantic flow effectively.

Overall, current fusion strategies insufficiently model multimodal heterogeneity, structure, and hierarchy, motivating the need for a more flexible and interpretable fusion mechanism that captures both fine-grained and coarse-grained inter-modal relations. Recent studies have further explored dynamic weighting strategies to adaptively adjust modality importance (Feng et al., 2024). In contrast, our FEMF follows a widely accepted assumption that the text modality provides the most reliable and fine-grained emotional cues, thereby serving as the dominant modality to guide multimodal fusion (Xie et al., 2024; Shi et al., 2025).

## 3 Methods

### 3.1 Problem definition

In the MSA task, the input data consists of three modalities: text($t$), audio($a$), and vision($v$). A multimodal utterance is represented as a triplet $(x_t, x_a, x_v)$, where $x_m \in \mathbb{R}^{T_m \times d_m}$ denotes the sequence of modality $m \in \{t, a, v\}$; $T_m$ is the sequence length, and $d_m$ is the feature dimension. Audio, text, and visual features are extracted by pre-trained wav2vec-large (Schneider et al., 2019), DeBERTa-large (He et al., 2021), and MA-Net (Zhao et al., 2021) (after MTCNN (Zhang et al., 2016) face alignment), respectively. The prediction target is the sentiment score $\hat{y} \in [-3, 3]$, where values greater than, equal to, and less than 0 correspond to positive, neutral, and negative sentiments, respectively.

### 3.2 Feature disentanglement module

Each modality-specific representation $x_m$ obtained from pre-trained encoders is further decomposed into two complementary components. The modality-shared representation encodes information within a common latent space while enforcing distributional similarity across modalities, thereby mitigating the heterogeneity gap and enabling more effective fusion. In contrast, the modality-private representation preserves modality-specific characteristics that are essential for capturing unique information cues. Together, these components are jointly optimized within a unified framework to balance cross-modal consistency and modality distinctiveness.

As shown in Fig. 1, We use a shared multimodal encoder $E^s$ and three separate unimodal encoders $E^u_m$, where $m \in \{a, t, v\}$, to learn the mapping representations, defined as follows:

$$x^s_m = E^s\big(x_m\big), \quad x^u_m = E^u_m\big(x_m\big) \tag{1}$$

Shared multimodal encoder $E^s$ and three separate unimodal encoders $E^u_m$ are composed of linear layers.

To reduce the discrepancy among the shared representations of the three modalities, we employ cosine similarity to quantify the differences, denoted as $L_{sim}$, between their respective distributions. Accordingly, within the shared representation space of all modalities, our objective is to minimize the following function:

$$\mathcal{L}_{\text{sim}} = \frac{1}{3} \sum_{(i,j)\in\mathcal{P}} \Big[1 - \frac{\langle x^s_{m_i}, x^s_{m_j} \rangle}{\|x^s_{m_i}\| \, \|x^s_{m_j}\|}\Big], \quad \mathcal{P} = \{(t,a),(t,v),(a,v)\} \tag{2}$$

Additionally, an orthogonal loss is introduced to guarantee that the modality-shared and modality-private representations focus on distinct attributes of the input data. This non-redundancy is enforced by applying an orthogonality constraint between these two types of representations. The corresponding orthogonality loss is computed as follows:

$$\mathcal{L}_{\text{orth}} = \frac{1}{3} \sum_{m\in\{a,t,v\}} \Big\langle x^s_m, \, x^u_m \Big\rangle \tag{3}$$

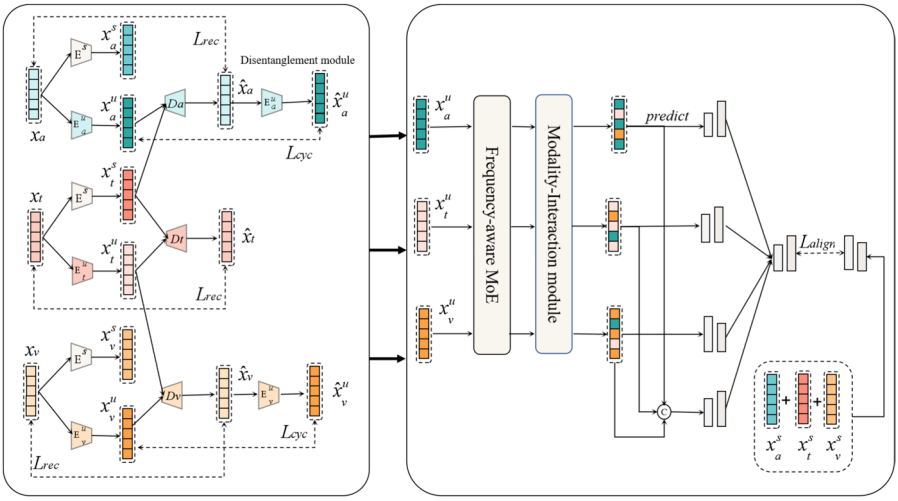


**Fig. 1** Overview of the proposed FEMF framework. The framework follows a pipeline of disentanglement, frequency-aware MoE, Modality-Interaction Module and Hierarchical Predictions Module

$\langle , \rangle$represents orthogonal operation. Although incorporating $L_{orth}$ helps enforce separation between shared and private representations, there remains a risk that modality-specific encoders could learn trivial or uninformative solutions. To address this concern, we integrate a modality reconstruction loss, which encourages the hidden representations to retain essential information from each modality and accurately reflect their respective characteristics.

Since the text modality has the highest confidence level among the three modalities, we concatenate the modality-specific representation of each modality $x_m^u$ and the modality-shared representation of the text modality $x_t^s$ as the input to the single decoder $D_m$, where $m \in \{a, t, v\}$, to reconstruct the corresponding representation. The discrepancy between the original multimodal representation and its reconstructed counterpart is quantified as follows:

$$\hat{x}_m = D_m\left(x_t^s,\, x_m^u\right) \tag{4}$$

$$\mathcal{L}_{\text{recon}} = \frac{1}{3} \sum_{m \in \{a,t,v\}} \text{MSE}\left(\hat{x}_m,\, x_m\right) \tag{5}$$

To ensure that the reconstructed modal representation retains the emotional semantic information of the original modality, we introduce an emotional consistency constraint mechanism to encourage the original representation and its reconstructed version to remain consistent in terms of emotional semantics. Inspired by the concept of cyclic consistency in the field of generative models (Wang et al., 2025b), we employ a cyclic consistency loss $\mathcal{L}_{cyc}$ to enhance the quality and robustness of representation reconstruction.

Specifically, the reconstructed modal representation is fed into its corresponding dedicated encoder for re-encoding, thereby regressing to its original modal unique representation. The difference between the reconstructed representation and the unique representation is defined as the cyclic consistency loss, as shown in the following formula:

$$\mathcal{L}_{cyc} = \sum_{m \in \{a,v\}} \|E_m^u(\hat{x}_m) - x_m^u\|_2^2 \tag{6}$$

Among them, $\hat{x}_m$ denotes the reconstructed modal representation, $E_m^u$ denotes the private encoder of modality *m*, and $x_m^u$ denotes the private representation of modality *m*. This loss term effectively ensures that the reconstruction path remains semantically consistent with the original modality, thereby enhancing the ability of the model to retain emotional semantics.Furthermore, since the reconstruction of the text modality is performed only within this modality, we did not constrain the text modality with this loss.

Finally, we integrate the above constraints to construct the overall loss function of the module, which is defined as follows:

$$\mathcal{L}_{\text{dis}} = \mathcal{L}_{\text{sim}} + \mathcal{L}_{\text{orth}} + \mathcal{L}_{\text{recon}} + \mathcal{L}_{\text{cyc}} \tag{7}$$

This work uses the private features of each modality as the main input for subsequent modules. Shared features are primarily used for alignment and supervision, while private features contain more modality-specific information such as voice intonation, visual details, or language style differences. Using private features as input for subsequent modules enhances complementary expression between modalities, improves the model's ability to model fine-grained emotional differences, and thereby enhances the final sentiment analysis performance and robustness.

### 3.3 Frequency-aware MoE

To further enhance the modeling capabilities of each modality representation, we decomposes each modality representation into low-frequency and high-frequency components, as shown in Fig. 2. This design is inspired by the observation that emotional semantics may simultaneously exhibit global (low-frequency) trends and sudden local changes (high-frequency cues).

**Discrete Fourier Transform (DFT)** Formally, we apply the real-valued Discrete Fourier Transform (DFT) along the temporal dimension to project the signal into the frequency domain.The Discrete Fourier Transform (DFT) serves as a fundamental tool in digital signal processing (DSP), enabling the transformation of discrete-time signals from the time domain into the frequency domain. We denote the DFT operation as a linear mapping:$\mathcal{F} : \mathbb{R}^N \to \mathbb{C}^N$, with its inverse–known as the Inverse Discrete Fourier Transform (IDFT)–represented as: $\mathcal{F}^{-1} : \mathbb{C}^N \to \mathbb{R}^N$, Applying $\mathcal{F}$ to a real-valued sequence $X \in \mathbb{R}^N$ corresponds to multiplying *X* by a DFT matrix $\mathcal{F}$. Each row of this matrix is a Fourier basis vector $f_j \in \mathbb{R}^N$, defined as:$\mathrm{f}_j = \left[e^{2\pi i(j-1)\cdot 0}, e^{2\pi i(j-1)\cdot 1}, \ldots, e^{2\pi i(j-1)(N-1)}\right]^T / \sqrt{N}$, where *i* is the imaginary unit, and $j \in \{1, ..., N\}$. Let the frequency spectrum of the signal be: $\mathrm{s}_x = \mathcal{F}(\mathrm{x}) \in \mathbb{C}^N$. We divide the spectrum into low- and high-frequency components: Low-frequency part $\mathrm{s}_x^{\mathrm{lfc}} \in \mathbb{C}^c$ and High-frequency part $\mathrm{s}_x^{\mathrm{hfc}} \in \mathbb{C}^{N-c}$. This decomposition is implemented via a hyperparameter *c* from our FourierLayer, and is applied to each modality's private representation independently:

$$x_m^{\text{low}},\, x_m^{\text{high}} = \text{FourierLayer}\left(x_m^u\right),\ m \in \{a, v, t\} \tag{8}$$

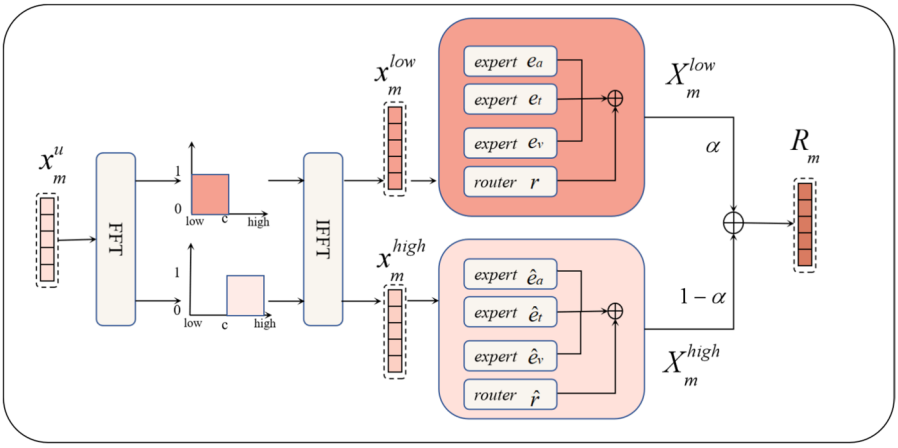


**Fig. 2** Overview of the Frequency-Aware MoE module. Each modality feature is decomposed into low- and high-frequency components via DFT, processed by frequency-specific experts, and fused adaptively through a soft router guided by the text modality

**Modality experts** Each frequency band obtained from the previous module is processed by a dedicated frequency-specific modality expert. Each expert comprises a stack of $Q$ Transformer Encoder blocks–each containing a multi-head self-attention layer and a feed-forward network, followed by residual connections and layer normalization. Unless otherwise specified, $Q$=4.

$$\hat{y}_m^f = \mathrm{FC}_m^f\left(\mathrm{Transformer}_m^f\left(\mathrm{x}_m^f\right)\right), f \in \{\mathrm{low}, \mathrm{high}\},\ m \in \{a, v, t\} \tag{9}$$

The training objective for each expert is to minimize the mean squared error (MSE) between its prediction and the ground truth:

$$\mathcal{L}_{train}^f = \mathrm{MSE}\left(y, \hat{y}_m^f\right), f \in \{\mathrm{low}, \mathrm{high}\},\ m \in \{a, v, t\} \tag{10}$$

After training, the parameters of all modality-frequency experts are frozen and reused in the downstream model. This design ensures that each expert captures discriminative patterns unique to either the stable (low-frequency) or dynamic (high-frequency) components of modality-specific sentiment signals.Through this process, each modality–frequency expert develops the ability to reinterpret input representations from its own modality viewpoint and integrate complementary cues learned from other modalities, forming a unified cross-modal understanding. Given the low-frequency component $x_m^{low}$ of the $m$-th modality, we feed it into every expert $e_i(\cdot)$ to obtain modality expert-specific embeddings as follows:

$$x_i^m = e_i\left(x_m^{low}\right) \tag{11}$$

Here, $e_i(\cdot)$ denotes the expert corresponding to the $i$-th modality (audio, visual, or text). This process allows each expert to reinterpret the input feature from its own modality perspective, producing a set of cross-expert representations $x_i^m$ that capture complementary knowledge from multiple modalities.

To further enhance representation expressiveness, we introduce a lightweight Soft Router, which is a two-layer Multilayer Perceptron (MLP), we treat the text modality as a high-confidence semantic anchor due to its empirical reliability in sentiment understanding. the router takes both the current modality feature and the text feature as input to estimate the contribution scores over all experts, achieving dynamic optimization and rebalancing of contributions from different experts (using low-frequency signals as an example):

$$S_m = [s_a^m,\ s_v^m,\ s_t^m] = \mathrm{MLP}(x_m^{low}, x_t^{low}),\ m \in \{a, v, t\} \tag{12}$$

$$w_i^m = \mathrm{softmax}(s_i^m),\ m \in \{a, v, t\} \tag{13}$$

$$X_m^{low} = \sum_{i \in \{a,v,t\}} w_i^m \cdot x_i^m,\ m \in \{a, v, t\} \tag{14}$$

which $w_i^m$ is the weight of the output representing of each expert. $X_m^{low}$ is the characteristic of low-frequency information of mode *m*.

Similarly, we can obtain the representation of the high-frequency information of mode *m* after expert processing $X_m^{high}$.

Subsequently, we introduce a hyperparameter $\alpha \in [0, 1]$ to control the contribution of high-frequency and low-frequency components, and fuse them through a weighted summation as follows:

$$R_m = \alpha \cdot X_m^{low} + (1 - \alpha) \cdot X_m^{high} \tag{15}$$

Unlike conventional feature partitioning that splits latent dimensions arbitrarily, our DFT-based decomposition performs signal-level separation grounded in the temporal–spectral characteristics of multimodal data. Through the Discrete Fourier Transform (DFT), each modality is divided into low-frequency components representing stable contextual trends and high-frequency components capturing transient emotional variations. This principled spectral separation enables frequency-specific experts to model complementary dynamics guided by interpretable frequency semantics, rather than arbitrary feature subsets. The resulting frequency-aware representations are then fed into cross-modal attention and hierarchical prediction layers to enhance semantic alignment and decision robustness.

### 3.4 Modality-interaction module

After generating modality-specific representations refined by the frequency-aware expert mechanism. We introduce a multi-head attention-based fusion module as shown in Fig. 3 that explicitly models the interactions between different modalities in a pairwise and sequential manner. This mechanism iteratively aligns each target modality with the complementary information from the remaining modalities using stacked bidirectional multi-head attention layers.

For each modality, the fusion process is performed in two stages. First, the target modality attends to another modality using its own features as queries and the other modality as key–value pairs. The resulting features are then fused with the third modality in the same manner. This staged attention mechanism enables each modality to progressively incorpo-

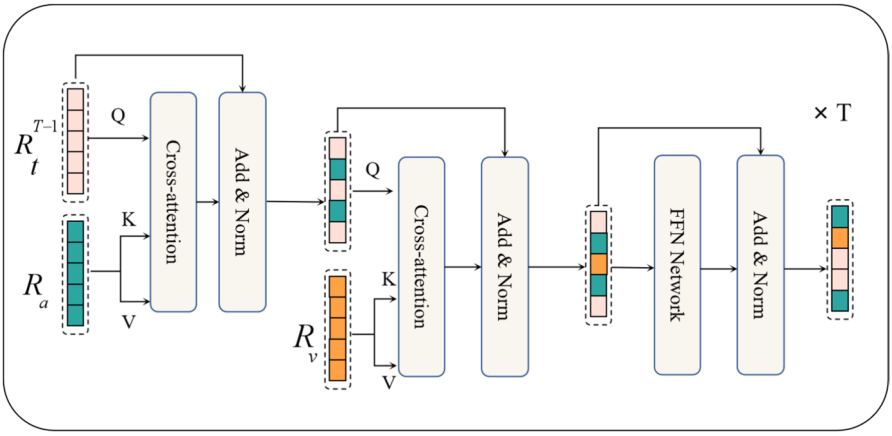


**Fig. 3** Modality-Interaction Module. The module is symmetrically three-branched, with one modality per branch as a query, illustrated in the figure with a text modality

rate semantic cues from others. The process is applied symmetrically across all modalities and repeated over $T$ stacked layers to facilitate deeper and more structured cross-modal interactions. Unless otherwise specified, $T$ is treated as a tunable hyperparameter.

Taking the textual modality as an example, let $R_t^{j-1}$ denote the textual representations at layer $j-1$, and let $R_a,\ R_v$ be the audio and visual modality features. The fusion process at layer $j$ includes:

**Text-audio fusion** The text features are first used as query vectors, while audio features are used as keys and values in a multi-head cross-attention block, modeling text-audio interaction.

$$A_{ta}^{j} = \text{Softmax}\left(\frac{R_t^{j-1} W_Q^{ta} (R_a W_K^{ta})^\top}{\sqrt{d}}\right) \cdot (R_a W_V^{ta}) \tag{16}$$

$$R_{ta}^{j} = \text{LayerNorm}(R_t^{j-1} + A_{ta}^{j}) \tag{17}$$

where $W_Q^{ta}, W_K^{ta}, W_V^{ta}$ are projection matrices.

Regarding the fusion of visual aspects, similar to the previous step, the fused text-audio features are used as queries, while the visual features are used as keys and values in the second cross-attention block to model text-visual interactions, and finally the fused features $R_{tav}^{j}$ are obtained.

**Feedforward refinement** A feed-forward network (FFN) with residual connection and layer normalization is applied to refine the fused features and enable deeper representations.

$$\text{MLP}(R_{tav}^{j}) = \text{ReLU}(R_{tav}^{j} W_1 + b_1) W_2 + b_2 \tag{18}$$

$$R_t^{j} = \text{LayerNorm}(R_{tav}^{j} + \text{MLP}(R_{tav}^{j})) \tag{19}$$

where $W_1, W_2$ are projection matrices. $b_1, b_2$ are bias parameters.

### 3.5 Hierarchical predictions module

To further enhance the discriminative capacity of each modality and the robustness of final decisions, we design a hierarchical prediction module that integrates both modality-specific and multimodal fusion predictions.

Specifically, for each refined modality representation $R_m$, where $m \in \{a, v, t\}$ , we apply a two-layer feed-forward network with residual connections to obtain the high-level representation, then, a modality-specific prediction is obtained by applying a sigmoid-activated projection:

$$h_m = R_m + \mathrm{FFN}(R_m), \quad \hat{y}_m = \mathrm{FC}(h_m) \tag{20}$$

To leverage the synergy among different modalities, we concatenate the refined modality representations to form a joint feature, and compute a fusion-based prediction:

$$R_{\mathrm{joint}} = [R_a; R_v; R_t], \quad \hat{y}_{joint} = \mathrm{FC}(R_{joint}) \tag{21}$$

To achieve adaptive fusion of hierarchical predictions, we use learnable fusion to calculate weights. Specifically, we employ a lightweight multi-layer perceptron (MLP) to estimate the importance of each prediction head, including three modality-specific predictions and a joint fusion prediction.Given the four prediction vectors, we first concatenate them into a combined vector, and this vector is passed through an MLP followed by a softmax layer to produce a set of normalized weights $\boldsymbol{\alpha} \in \mathbb{R}^4$:

$$z = [\hat{y}_a; \hat{y}_v; \hat{y}_t; \hat{y}_{\mathrm{joint}}], \quad \alpha = \mathrm{Softmax}(\mathrm{MLP}(z)) \tag{22}$$

The final prediction $\hat{y}$ is obtained via a weighted combination of all individual predictions:

$$\hat{y} = \sum_{i=1}^{4} \alpha_i \hat{y}^{(i)}, \quad \hat{y}^{(i)} \in \{\hat{y}_a, \hat{y}_v, \hat{y}_t, \hat{y}_{joint}\} \tag{23}$$

This design allows the model to adaptively emphasize more reliable modality-specific or joint outputs under different input conditions.

In addition, we aggregate the modality-shared representations obtained from the disentanglement module–which are not directly used in subsequent prediction steps–and use them to produce an auxiliary prediction. A consistency constraint is then applied between this auxiliary prediction and the final output.

$$\hat{y}_{shared} = FC(x_a^s + x_t^s + x_v^s) \tag{24}$$

$$L_{align} = \mathrm{MSE}(\hat{y}_{shared}, \hat{y}) \tag{25}$$

This constraint encourages the shared representations to preserve task-relevant semantics and align their predictive behavior with the final decision, thereby improving the robustness and interpretability of the shared feature space.

### 3.6 Overall learning objective

The final MSA output loss $L_{task}$ is defined as:

$$\mathcal{L}_{task} = \frac{1}{N}\sum_{n=1}^{N}(\hat{y}_n - y_n)^2 \tag{26}$$

where $y_n$ is the MSA label, $n$ is the number of samples. And the total modal experts training loss $L_{train}$ is defined as:

$$\mathcal{L}_{train} = \sum_{m\in\{a,v,t\}}\sum_{f\in\{\text{low,high}\}} \text{MSE}\left(y, \hat{y}_m^f\right) \tag{27}$$

The proposed framework integrates the decoupling loss $L_{dis}$, the modal experts training loss $L_{train}$, the shared consistency loss $L_{align}$ and the total MSA task loss $L_{task}$ to form the overall learning objective:

$$\mathcal{L}_{total} = \mathcal{L}_{dis} + \mathcal{L}_{train} + \mathcal{L}_{align} + \mathcal{L}_{task} \tag{28}$$

## 4 Experiments

### 4.1 Datasets

We evaluate the proposed model on two widely used benchmark datasets: CMU-MOSI (Zadeh et al., 2016) and CMU-MOSEI (Zadeh et al., 2018).

**CMU-MOSI** This dataset contains 2,199 monologue video segments of movie reviews, each accompanied by audio and visual streams sampled at 12.5 Hz and 15 Hz, respectively. The dataset is split into 1,284 training samples, 229 validation samples, and 686 testing samples.

**CMU-MOSEI** As a larger-scale extension, this corpus consists of 22,856 video clips collected from YouTube, with audio features sampled at 20 Hz and visual features at 15 Hz. It includes 16,326 training samples, 1,871 validation samples, and 4,659 testing samples. For both datasets, sentiment annotations range from $-3$ to $+3$, representing a fine-grained continuum from strongly negative to strongly positive emotions.

### 4.2 Evaluation metrics

Following prior studies (Wang et al., 2025b; Yang et al., 2023b), we adopt a comprehensive set of evaluation metrics to assess model performance, including binary accuracy (Acc-2),

F1 score, Pearson correlation coefficient (Corr) between predicted and ground-truth sentiment scores, and Mean Absolute Error (MAE). For binary accuracy, we report both negative/non-negative and negative/positive accuracies using the "$-/-$" notation, where the left value represents the former and the right value denotes the latter, ensuring a fair and detailed comparison across studies.

### 4.3 Implementation details

In this study, we follow the experimental settings of prior works (Wang et al., 2025b; Xu et al., 2024). All experiments are implemented in PyTorch and conducted on an NVIDIA RTX 3090 GPU with 24 GB memory. The model is trained with a batch size of 64 and a hidden dimension of 256, consistent with the feature dimension of both modality-specific and cross-modal attention embeddings. We employ the Adam optimizer with a learning rate of 0.0001 and run each experiment three times, reporting the average performance. For fair comparison, all pre-trained encoders in FEMF are identical to those used in MoMKE: DeBERTa-large for text, wav2vec-large for audio, and MA-Net for visual features, ensuring complete consistency across modalities.

### 4.4 Baselines

Self-MM (Yu et al., 2021b) performs joint unimodal and multimodal learning via a self-supervised label module. MMIM (Han et al., 2021) enhances fusion by maximizing mutual information across modalities. CubeMLP (Sun et al., 2022) models cross-modal interactions using a cube-structured MLP. ConFEDE (Yang et al., 2023b) applies contrastive disentanglement and knowledge distillation for robust fusion. ALMT (Zhang et al., 2023) introduces language-guided adaptive modules to filter irrelevant signals. TMSON (Xie et al., 2024) models modality uncertainty via Gaussian fusion with ordinal sentiment space. HyDiscGAN (Wu et al., 2024) employs hybrid discriminators for global–local alignment. MoMKE (Xu et al., 2024) integrates unimodal and joint features through a soft routing strategy. DEVA (Wu et al., 2025) generates textual emotional descriptions from audio–visual cues for better semantic reasoning. MFMB-Net (Tao et al., 2025) adopts macro–micro fusion and modality reconstruction to improve robustness.

### 4.5 Performance comparison

As shown in Table 1, the proposed model achieves competitive or superior results across most evaluation metrics on both CMU-MOSI and CMU-MOSEI. Compared with prior state-of-the-art approaches, FEMF exhibits notable improvements in Pearson correlation, classification accuracy, and F1 score, reflecting a stronger alignment between predicted and ground-truth sentiment scores. These gains indicate that FEMF better captures nuanced sentiment dynamics within multimodal inputs.

Under both the negative/non-negative and negative/positive binary protocols, our model consistently achieves higher accuracies and F1 scores, demonstrating stronger discriminative capability and robustness across datasets. Compared with mutual-information or generative models such as Self-MM, MMIM, and HyDiscGAN, FEMF achieves more consistent improvements across both correlation and classification metrics. While confidence-aware

**Table 1** Performance comparison on CMU-MOSI and CMU-MOSEI datasets. "Acc-2" and "F1" are reported as neg./non-neg. and neg./pos. respectively. Our model use the same feature extractors as MoMKE, including DeBERTa-large (text), wav2vec-large (audio), and MA-Net (visual)

| Model | CMU-MOSI | | | | CMU-MOSEI | | | |
|---|---|---|---|---|---|---|---|---|
| | MAE | Corr | Acc-2 | F1 | MAE | Corr | Acc-2 | F1 |
| Self-MM | 0.713 | 0.798 | 84.0/86.0 | 84.4/86.0 | 0.539 | 0.753 | 83.8/85.2 | 83.7/85.1 |
| MMIM | 0.712 | 0.790 | 83.34/85.39 | 83.43/85.41 | 0.536 | 0.764 | 82.57/85.01 | 82.41/85.13 |
| CubeMLP | 0.755 | 0.772 | 80.76/82.32 | 81.77/84.23 | 0.537 | 0.761 | 82.36/85.23 | 82.61/85.04 |
| ConFEDE | 0.742 | 0.784 | 84.17/85.52 | 84.13/85.52 | **0.522** | 0.780 | 81.65/85.82 | 82.17/85.83 |
| ALMT | 0.712 | 0.793 | 83.97/85.82 | 84.05/85.86 | 0.530 | 0.774 | 81.54/85.99 | 81.05/86.05 |
| TMSON | **0.687** | 0.809 | 85.40/87.2 | **85.40**/87.2 | 0.526 | 0.766 | 85.2/86.4 | 85.3/86.2 |
| HyDiscGAN | 0.749 | 0.782 | 84.1/86.7 | 83.7/86.3 | 0.533 | 0.761 | 81.9/86.3 | 82.1/86.2 |
| MoMKE* | 0.798 | 0.801 | 85.27/87.96 | 85.22/87.97 | 0.556 | 0.827 | 84.97/87.29 | 84.91/87.19 |
| DEVA | 0.73 | 0.787 | 84.4/86.29 | 84.48/86.3 | 0.541 | 0.769 | 83.26/86.13 | 82.93/86.21 |
| MFMB-Net | 0.709 | 0.798 | 82.7/85.70 | 83.2/86.00 | 0.532 | 0.758 | 84.7/85.10 | 85.0/85.10 |
| **Ours** | 0.758 | **0.828** | **85.48/88.37** | 85.32/**88.36** | 0.553 | **0.831** | **85.23/87.6** | **85.41/87.54** |

methods like ConFEDE and TMSON enhance robustness via uncertainty modeling and ordinal representations, FEMF further strengthens representation learning through the joint modeling of shared and modality-specific features under hierarchical supervision. Even relative to advanced semantic alignment or routing frameworks such as DEVA and MoMKE, FEMF maintains favorable overall performance, confirming its effectiveness in balancing fine-grained regression accuracy with robust classification in multimodal sentiment analysis.

## 4.6 Ablation study

### 4.6.1 Effects of different components

To evaluate the contribution of each component, we perform ablation studies by individually removing the Disentanglement Module (DTM), Frequency-Aware MoE (FAE), Modality-Interaction Module (MIM), Hierarchical Prediction Module (HPM), and five auxiliary objectives ($L_{\text{align}}$, $L_{\text{rec}}$, $L_{\text{cyc}}$, $L_{\text{orth}}$, $L_{\text{sim}}$), while keeping the remaining architecture unchanged. The results are summarized in Tables 2 and 3.

Removing the **DTM** causes a clear decline across all metrics, confirming its role in generating informative and disentangled modality-specific representations. Excluding the **FAE**—thus omitting frequency decomposition and expert selection—leads to higher errors and weaker consistency, validating the benefit of frequency-aware modeling for capturing stable and transient emotional cues. Eliminating the **MIM** substantially reduces accuracy and correlation, showing that explicit inter-modal interaction modeling is essential for robust fusion. Removing the **HPM** and using single-stage prediction also degrades performance, highlighting the value of hierarchical supervision for complementary unimodal–multimodal learning and improved generalization.

Regarding auxiliary objectives, removing $L_{\text{align}}$ weakens shared-representation consistency, while omitting $L_{\text{rec}}$ or $L_{\text{cyc}}$ moderately harms performance, underscoring their role in stable reconstruction and cross-modal coherence. Excluding $L_{\text{orth}}$ increases redundancy between shared and private spaces, and dropping $L_{\text{sim}}$ slightly reduces accuracy, indicating its contribution to high-level relational alignment.

**Table 2** Ablation study results on the CMU-MOSI dataset

| Model | MAE | Corr | Acc-2 | F1 |
|---|---|---|---|---|
| Ours | **0.758** | **0.828** | **85.48/88.37** | **85.32/88.36** |
| w/o DTM | 0.772 | 0.814 | 84.92/87.91 | 84.89/87.87 |
| w/o FAE | 0.802 | 0.817 | 85.09/88.01 | 85.06/88.00 |
| w/o MIM | 0.821 | 0.819 | 85.03/87.96 | 84.93/87.92 |
| w/o HPM | 0.779 | 0.828 | 84.94/88.11 | 84.90/88.07 |
| w/o $L_{align}$ | 0.787 | 0.827 | 85.39/88.27 | 85.21/88.26 |
| w/o $L_{rec}$ | 0.772 | 0.827 | 85.41/88.33 | 85.29/88.33 |
| w/o $L_{cyc}$ | 0.772 | 0.826 | 85.4/88.33 | 85.22/88.32 |
| w/o $L_{orth}$ | 0.786 | 0.826 | 85.38/88.25 | 85.29/88.25 |
| w/o $L_{sim}$ | 0.782 | 0.825 | 85.37/88.26 | 85.27/88.25 |

**Table 3** Ablation study results on the CMU-MOSEI dataset

| Model | MAE | Corr | Acc-2 | F1 |
|---|---|---|---|---|
| Ours | **0.553** | **0.831** | **85.23/87.6** | **85.41/87.54** |
| w/o DTM | 0.566 | 0.821 | 84.99/87.53 | 85.02/87.45 |
| w/o FAE | 0.557 | 0.828 | 84.81/87.41 | 84.87/87.38 |
| w/o MIM | 0.553 | 0.830 | 84.85/87.47 | 84.91/87.43 |
| w/o HPM | 0.556 | 0.829 | 84.93/87.46 | 84.89/87.42 |
| w/o $L_{align}$ | 0.555 | 0.828 | 84.96/87.51 | 84.98/87.41 |
| w/o $L_{rec}$ | 0.553 | 0.83 | 85.01/87.56 | 84.93/87.49 |
| w/o $L_{cyc}$ | 0.562 | 0.829 | 84.97/87.58 | 84.91/87.51 |
| w/o $L_{orth}$ | 0.557 | 0.83 | 84.9/87.46 | 84.88/87.39 |
| w/o $L_{sim}$ | 0.56 | 0.831 | 85.01/87.53 | 84.97/87.46 |

Overall, these findings confirm that both architectural components and auxiliary losses contribute uniquely and complementarily to the model's final performance, validating the necessity and effectiveness of the full FEMF design.

### 4.6.2 Effects of different modalities

As shown in Table 4, the textual modality delivers the strongest unimodal performance, significantly surpassing the visual and auditory counterparts, which underscores the rich semantic information conveyed by language in emotion understanding. In the bimodal setting, combining any two modalities consistently improves performance over unimodal baselines, reflecting the complementary nature of multimodal cues. Furthermore, the trimodal configuration achieves the best overall results on both benchmarks, confirming that our framework effectively leverages heterogeneous yet complementary information. The progressive improvements from unimodal to bimodal and finally to trimodal fusion demonstrate the robustness of our multimodal integration strategy in capturing cross-modal dependencies and enhancing fine-grained sentiment discrimination.

## 4.7 Parameter sensitivity analysis

As illustrated in Fig. 4, we conduct a sensitivity analysis on three key parameters–the number of layers in the modality interaction module, the fusion weight $\alpha$ in the frequency-aware

**Table 4** Performance comparison of different modality combinations on CMU-MOSI and CMU-MOSEI datasets. T = Text modality, A = Audio modality, V = Visual modality

| | CMU-MOSI | | | CMU-MOSEI | | |
|---|---|---|---|---|---|---|
| Model | MAE | Acc-2 | F1 | MAE | Acc-2 | F1 |
| T | 0.762 | 87.25 | 87.19 | 0.564 | 86.99 | 86.97 |
| A | 1.482 | 54.93 | 55.02 | 0.883 | 72.48 | 71.80 |
| V | 1.409 | 60.91 | 60.89 | 0.908 | 69.45 | 68.76 |
| T+A | 0.771 | 87.49 | 87.46 | 0.558 | 87.28 | 87.22 |
| T+V | 0.852 | 88.01 | 88.02 | 0.563 | 87.13 | 87.06 |
| A+V | 1.373 | 61.43 | 61.45 | 0.868 | 72.52 | 71.73 |
| T+A+V | **0.758** | **88.37** | **88.36** | **0.553** | **87.60** | **87.54** |

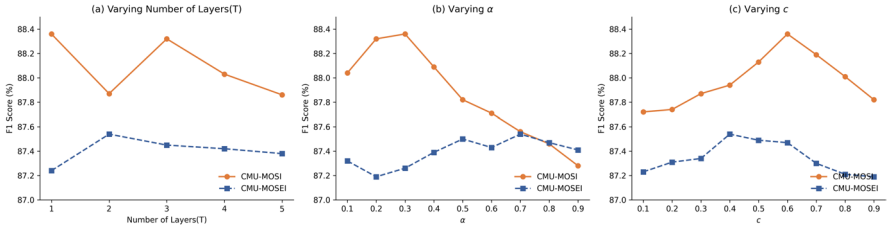


**Fig. 4** Parameter Sensitivity Analysis

Mixture-of-Experts (MoE) branch, and the frequency truncation coefficient $c$—to examine their impact on model performance across CMU-MOSI and CMU-MOSEI.

For the number of layers (Fig. 4(a)), CMU-MOSI achieves the best performance with a single layer, while deeper architectures lead to slight degradation, likely due to overfitting or redundant modeling. CMU-MOSEI, with its larger data scale, exhibits greater stability across layer depths. Varying the fusion coefficient $\alpha$ (Fig. 4(b)) within [0.1, 0.9] shows an optimal value around 0.3, where assigning moderate emphasis to high-frequency components yields stronger discriminative representations. Extreme values reduce performance, confirming the importance of balanced frequency fusion. Finally, the truncation coefficient $c$ (Fig. 4(c)), which determines the cutoff between low- and high-frequency components in the Fourier-based decomposition, peaks at $c = 0.6$ on CMU-MOSI and around $c = 0.4$ on CMU-MOSEI. This indicates that moderate retention of high-frequency information enhances sensitivity to fine-grained emotional cues, while larger and more diverse datasets remain less affected by the specific cutoff.

## 4.8 Modality interaction order analysis

To assess the impact of modality interaction order within each query-dominant branch of our modality-interaction module, we conduct an experiment exploring different sequential fusion paths among modalities, as illustrated in Fig. 5.

For the text-query branch, interacting with the audio modality first yields the best performance across all metrics, indicating that early fusion with temporally aligned acoustic cues refines semantic grounding before integrating higher-level visual information. In the audio-query branch, the optimal order is A→T→V, showing that contextualizing audio with

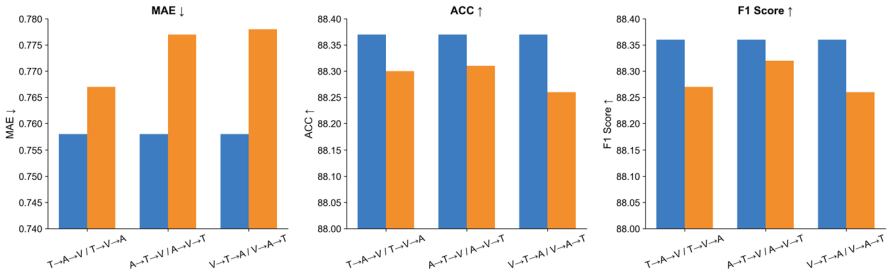


**Fig. 5** Modality Interaction Order Analysis on MOSI

linguistic information enhances sentiment interpretation more effectively than initiating with less aligned visual features. For the visual-query branch, the V→T→A sequence consistently outperforms its reverse, suggesting that incorporating textual context early helps disambiguate visual signals prior to adding acoustic nuances.

Overall, these findings reveal that the order of modality interaction plays a non-trivial role. The proposed multi-branch interaction module benefits from carefully designed fusion sequences, effectively capturing diverse and complementary cross-modal dependencies.

### 4.9 Robustness analysis under missing modalities

To evaluate model robustness under incomplete multimodal inputs, we conduct experiments where input features are randomly masked with missing rates ranging from 0% to 40%, simulating real-world conditions such as sensor failure or transmission noise.

As shown in Fig. 6, all models experience performance degradation as the missing rate increases, revealing the inherent difficulty of incomplete multimodal learning. However, our model exhibits a markedly slower decline in both F1 score and Pearson correlation compared with representative baselines, demonstrating stronger resilience to modality loss. This robustness stems from the disentangled representation learning framework, which jointly captures complementary information from shared and modality-specific spaces, as well as from the confidence-aware fusion and hierarchical prediction strategy that adaptively emphasizes reliable modalities.

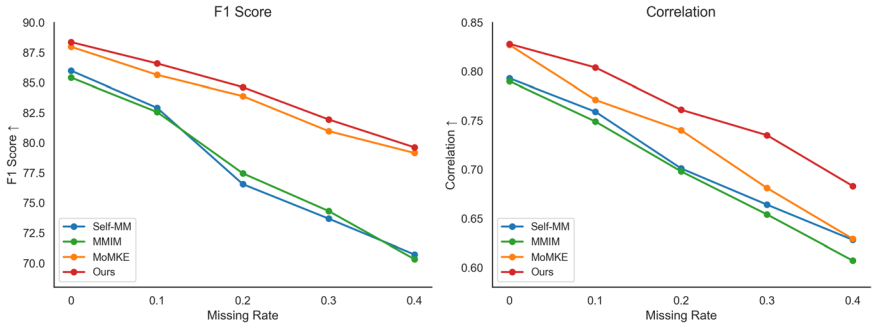


**Fig. 6** Robustness Analysis under Missing Modalities

Moreover, this robustness analysis also reflects the model's data generalization capability: FEMF maintains stable performance even with substantial modality absence, indicating strong adaptability under limited-data or few-shot conditions. The frequency-aware multi-stage fusion mechanism further contributes to cross-modal alignment and enhances resistance to noise or missing signals, underscoring the practicality of our approach in real-world multimodal scenarios.

### 4.10 Visualization analysis

To assess the effectiveness of our Fourier-aware Expert Modeling, we visualized the attention distributions from the high- and low-frequency branches across modalities (Fig. 7). The results show clear and modality-specific divergence: audio high-frequency components focus on prosodic shifts, while low-frequency ones capture the global tone; text high-frequency branches attend to emotionally charged words, whereas low-frequency ones reflect stable semantic structures; visual high-frequency responses highlight local edges and rapid facial changes, while low-frequency responses capture overall appearance and scene context. These complementary patterns confirm that the frequency-aware modeling enables adaptive and semantically meaningful attention allocation. To further clarify the semantic implication of the frequency decomposition, we note that high-frequency components capture transient and emotionally intense variations such as abrupt vocal or facial changes, while low-frequency components encode smoother and more stable emotional contours representing calm or sustained tones. This observation aligns with the intended design of the frequency-aware experts and supports the effectiveness of the proposed decomposition.

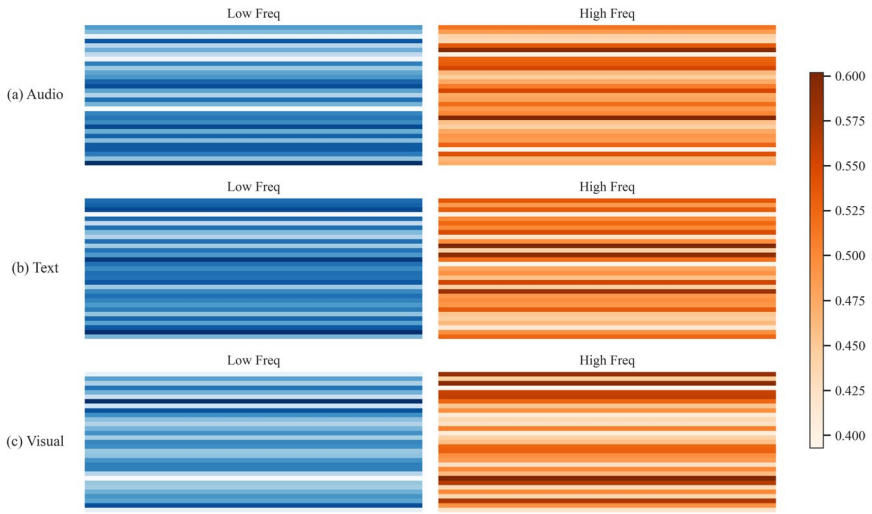


**Fig. 7** Visualization of attention distributions from the low- and high-frequency branches across audio, text, and visual modalities. High-frequency features emphasize abrupt and emotionally intense cues (e.g., sudden intonation or facial muscle changes), whereas low-frequency features capture stable and context-consistent patterns (e.g., calm tone or overall facial expression)

## 5 Conclusion

In this paper, we proposed FEMF, a Frequency-Aware Experts with Multi-Stage Fusion framework for multimodal sentiment analysis. By incorporating frequency decomposition via Discrete Fourier Transform and expert modeling for high- and low-frequency components, FEMF effectively captures both stable and transient emotional cues across modalities. The integration of shared–private disentanglement, modality-interaction modules, and hierarchical predictions further enhances robustness and interpretability. Extensive experiments on CMU-MOSI and CMU-MOSEI demonstrate consistent improvements over strong baselines, while ablation and robustness analyses confirm the contribution of each component and the overall resilience of the architecture.

For future work, we aim to extend FEMF in several directions: (1) exploring adaptive spectral decomposition methods–such as wavelet and graph Fourier transforms–to better model non-periodic or context-dependent modalities like text; (2) applying FEMF to broader multimodal affective datasets (e.g., CrisisMMD, MER2024) to assess cross-domain generalization; (3) integrating LLM-based semantic enrichment to enhance text-level affective reasoning; and (4) developing low-rank and parameter-efficient MoE variants for improved scalability. In addition, we plan to incorporate explicit dynamic weighting mechanisms among modalities in the future work to make the fusion process more adaptive and interpretable. Furthermore, we will investigate how frequency-aware and cross-modal semantic representations can help mitigate ethical and social biases inherent in pre-trained multimodal encoders, thereby promoting fairness in affective computing applications.

**Author Contributions** Y. L. : Conceptualization, Methodology, Validation, Writing. X.Z. : Supervision, Resources, Funding, Editing.

**Funding** This work was supported by the Science and Technology Innovation Key R&D Program of Chongqing (CSTB2024TIAD-STX0027), the National Natural Science Foundation of China (62472059), the Chongqing Talent Plan Project, China (CSTC2024YCJH-BGZXM0022), the Open Research Fund of Key Laboratory of Cyberspace Big Data Intelligent Security (Chongqing University of Posts and Telecommunications), Ministry of Education (CBDIS202403).

**Data Availability** No datasets were generated or analysed during the current study.

### Declarations

**Ethical Approval** Not applicable.

**Competing interests** The authors declare no competing interests.

## References

Ai, W., Zhang, F., Shou, Y., et al. (2025). Revisiting multimodal emotion recognition in conversation from the perspective of graph spectrum. *Proceedings of the AAAI Conference on Artificial Intelligence, 39*(11), 11418–11426. https://doi.org/10.1609/aaai.v39i11.33242

Cheng, J., Zhu, X., & Yang, Z. (2025). Tf-merc: Integrating time-frequency information for multimodal emotion recognition in conversation. In: Proceedings of the 2025 International Conference on Multimedia Retrieval. Association for Computing Machinery, New York, NY, USA, ICMR '25, p 126–134, https://doi.org/10.1145/3731715.3733447

Cheung, M., Shi, J., Wright, O., et al. (2020). Graph signal processing and deep learning: Convolution, pooling, and topology. *IEEE Signal Processing Magazine, 37*(6), 139–149. https://doi.org/10.1109/MSP.2020.3014594

Feng, X., Lin, Y., He, L., et al. (2024). Knowledge-guided dynamic modality attention fusion framework for multimodal sentiment analysis. In: Al-Onaizan, Y., Bansal, M., & Chen, Y.N. (eds) Findings of the Association for Computational Linguistics: EMNLP 2024. Association for Computational Linguistics, Miami, Florida, USA, pp 14755–14766, https://doi.org/10.18653/v1/2024.findings-emnlp.865

Han, W., Chen, H., & Poria, S. (2021). Improving multimodal fusion with hierarchical mutual information maximization for multimodal sentiment analysis. In: Moens, M.F., Huang, X., Specia, L., et al. (eds) Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, pp 9180–9192, https://doi.org/10.18653/v1/2021.emnlp-main.723

Hazarika, D., Zimmermann, R., & Poria, S. (2020). Misa: Modality-invariant and -specific representations for multimodal sentiment analysis. In: Proceedings of the 28th ACM International Conference on Multimedia. Association for Computing Machinery, New York, NY, USA, MM '20, p 1122–1131, https://doi.org/10.1145/3394171.3413678

He, P., Liu, X., Gao, J., et al. (2021). Deberta: Decoding-enhanced bert with disentangled attention. In: International Conference on Learning Representations, https://openreview.net/forum?id=XPZIaotutsD

Hou, M., Tang, J., Zhang, J., et al. (2019). Deep multimodal multilinear fusion with high-order polynomial pooling. *Advances in Neural Information Processing Systems, 32*. https://doi.org/10.5555/3454287.3455376

Huang, J., Tao, J., Liu, B., et al. (2020). Multimodal transformer fusion for continuous emotion recognition. In: ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp 3507–3511, https://doi.org/10.1109/ICASSP40776.2020.9053762

Lepikhin, D., Lee, H., Xu, Y., et al. (2021). Gshard: Scaling giant models with conditional computation and automatic sharding. In: 9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021, https://arxiv.org/abs/2006.16668

Li, Y., Wang, Y., & Cui, Z. (2023). Decoupled multimodal distilling for emotion recognition. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp 6631–6640, https://doi.org/10.1109/CVPR52729.2023.00641

Li, M., Yang, D., Zhao, X., et al. (2024a). Correlation-decoupled knowledge distillation for multimodal sentiment analysis with incomplete modalities. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 12458–12468, https://doi.org/10.1109/CVPR52733.2024.01184

Li, Y., Ding, H., Lin, Y., et al. (2024). Multi-level textual-visual alignment and fusion network for multimodal aspect-based sentiment analysis. *Artificial Intelligence Review, 57*, 78. https://doi.org/10.1007/s10462-023-10685-z

Li, Y., Liu, A., & Lu, Y. (2025). Multi-level language interaction transformer for multimodal sentiment analysis. *Journal of Intelligent Information Systems, 63*(3), 945–964. https://doi.org/10.1007/s10844-025-00923-x

Li, Y., Zhu, R., & Li, W. (2025). Cormult: A semi-supervised modality correlation-aware multimodal transformer for sentiment analysis. *IEEE Transactions on Affective Computing, 16*(3), 2321–2333. https://doi.org/10.1109/TAFFC.2025.3559866

Li, Z., Tang, F., Zhao, M., et al. (2022). Emocaps: Emotion capsule based model for conversational emotion recognition. In: Findings of the Association for Computational Linguistics: ACL 2022, https://doi.org/10.18653/v1/2022.findings-acl.126

Liu, Z., Braytee, A., Anaissi, A., et al. (2024). Ensemble pretrained models for multimodal sentiment analysis using textual and video data fusion. In: Companion Proceedings of the ACM Web Conference 2024. Association for Computing Machinery, New York, NY, USA, WWW '24, p 1841–1848, https://doi.org/10.1145/3589335.3651971

Mustafa, B., Riquelme, C., Puigcerver, J., et al. (2022). Multimodal contrastive learning with limoe: the language-image mixture of experts. In: Proceedings of the 36th International Conference on Neural Information Processing Systems, pp 9564–9576, https://doi.org/10.5555/3600270.3600965

Ong, R.K., & Khong, A.W.H. (2025). Spectrum-based modality representation fusion graph convolutional network for multimodal recommendation. In: Proceedings of the Eighteenth ACM International Conference on Web Search and Data Mining. Association for Computing Machinery, New York, NY, USA, WSDM '25, p 773–781, https://doi.org/10.1145/3701551.3703561

Paraskevopoulos, G., Georgiou, E., & Potamianos, A. (2022). Mmlatch: Bottom-up top-down fusion for multimodal sentiment analysis. In: ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, pp 4573–4577, https://doi.org/10.1109/ICASSP43922.2022.9746418

Schneider, S., Baevski, A., Collobert, R., et al. (2019). wav2vec: Unsupervised pre-training for speech recognition. In: Interspeech, https://doi.org/10.21437/Interspeech.2019-1873

Shi, P., Hu, M., Nakagawa, S., et al. (2025). Text-guided reconstruction network for sentiment analysis with uncertain missing modalities. *IEEE Transactions on Affective Computing, 16*(3), 1825–1838. https://doi.org/10.1109/TAFFC.2025.3541743

Shin, Y., Choi, J., Wi, H., et al. (2024). An attentive inductive bias for sequential recommendation beyond the self-attention. *Proceedings of the AAAI Conference on Artificial Intelligence, 38*(8), 8984–8992. https://doi.org/10.1609/aaai.v38i8.28747

Sun, H., Wang, H., Liu, J., et al. (2022). Cubemlp: An mlp-based model for multimodal sentiment analysis and depression estimation. In: Proceedings of the 30th ACM International Conference on Multimedia, pp 3722–3729, https://doi.org/10.1145/3503161.3548025

Sun, Z., Sarma, P., Sethares, W., et al. (2020). Learning relationships between text, audio, and video via deep canonical correlation for multimodal language analysis. *Proceedings of the AAAI Conference on Artificial Intelligence, 34*(05), 8992–8999. https://doi.org/10.1609/aaai.v34i05.6431

Tao, C., Li, J., Zang, T., et al. (2025). A multi-focus-driven multi-branch network for robust multimodal sentiment analysis. In: Proceedings of the AAAI Conference on Artificial Intelligence, pp 1547–1555, https://doi.org/10.1609/aaai.v39i2.32146

Wang, H., Cao, J., Liu, J., et al. (2025). A method for multimodal sentiment analysis: adaptive interaction and multi-scale fusion. *Journal of Intelligent Information Systems, 63*, 1667–1686. https://doi.org/10.1007/s10844-025-00957-1

Wang, P., Zhou, Q., Wu, Y., et al. (2025b). Dlf: Disentangled-language-focused multimodal sentiment analysis. In: Proceedings of the AAAI Conference on Artificial Intelligence, pp 21180–21188, https://doi.org/10.1609/aaai.v39i20.35416

Wu, S., He, D., Wang, X., et al. (2025). Enriching multimodal sentiment analysis through textual emotional descriptions of visual-audio content. In: Proceedings of the AAAI Conference on Artificial Intelligence, AAAI '25, vol 39. Association for the Advancement of Artificial Intelligence, Palo Alto, CA, USA, pp 1601–1609, https://doi.org/10.1609/aaai.v39i2.32152

Wu, Z., Zhang, Q., Miao, D., et al. (2024). Hydiscgan: A hybrid distributed cgan for audio-visual privacy preservation in multimodal sentiment analysis. In: Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, https://doi.org/10.24963/ijcai.2024/724

Xie, Z., Yang, Y., Wang, J., et al. (2024). Trustworthy multimodal fusion for sentiment analysis in ordinal sentiment space. *IEEE Transactions on Circuits and Systems for Video Technology, 34*(8), 7657–7670. https://doi.org/10.1109/TCSVT.2024.3376564

Xu, W., Jiang, H., & Liang, X. (2024). Leveraging knowledge of modality experts for incomplete multimodal learning. In: Proceedings of the 32nd ACM International Conference on Multimedia, pp 438–446, https://doi.org/10.1145/3664647.3681683

Yang, D., Chen, Z., Wang, Y., et al. (2023a). Context de-confounded emotion recognition. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp 19005–19015, https://doi.org/10.1109/CVPR52729.2023.01822

Yang, J., Yu, Y., Niu, D., et al. (2023b). Confede: Contrastive feature decomposition for multimodal sentiment analysis. In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp 7617–7630, https://doi.org/10.18653/v1/2023.acl-long.421

Yang, K., Yang, D., Zhang, J., et al. (2023c). What2comm: Towards communication-efficient collaborative perception via feature decoupling. In: Proceedings of the 31st ACM International Conference on Multimedia. Association for Computing Machinery, New York, NY, USA, MM '23, p 7686–7695, https://doi.org/10.1145/3581783.3611699

Yu, W., Xu, H., Yuan, Z., et al. (2021). Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis. *Proceedings of the AAAI Conference on Artificial Intelligence, 35*, 10790–10797. https://doi.org/10.1609/aaai.v35i12.17289

Yu, W., Xu, H., Yuan, Z., et al. (2021b). Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis. In: Proceedings of the AAAI conference on Artificial Intelligence, pp 10790–10797, https://doi.org/10.1609/aaai.v35i12.17289

Zadeh, A., Zellers, R., Pincus, E., et al. (2016). Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages. *IEEE Intelligent Systems, 31*(6), 82–88. https://doi.org/10.1109/MIS.2016.94

Zadeh, A., Chen, M., Poria, S., et al. (2017a). Tensor fusion network for multimodal sentiment analysis. In: Palmer, M., Hwa, R., & Riedel, S. (eds) Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Copenhagen, Denmark, pp 1103–1114, https://doi.org/10.18653/v1/D17-1115

Zadeh, A., Chen, M., Poria, S., et al. (2017b). Tensor fusion network for multimodal sentiment analysis. In: Palmer, M., Hwa, R., & Riedel, S. (eds) Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Copenhagen, Denmark, pp 1103–1114, https://doi.org/10.18653/v1/D17-1115

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

Zhang, H., Wang, Y., Yin, G., et al. (2023). Learning language-guided adaptive hyper-modality representation for multimodal sentiment analysis. In: Bouamor, H., Pino, J., & Bali, K. (eds) Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Singapore, pp 756–767, https://doi.org/10.18653/v1/2023.emnlp-main.49

Zhang, K., Zhang, Z., Li, Z., et al. (2016). Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters, 23*(10), 1499–1503. https://doi.org/10.1109/LSP.2016.2603342

Zhang, Q., Zhu, X., Liu, Y., et al. (2019). Iris recognition based on adaptive optimization log-gabor filter and rbf neural network. In: Biometric Recognition. Springer International Publishing, Cham, pp 312–320, https://doi.org/10.1007/978-3-030-31456-9_3

Zhang, Q., Miao, D., Zhang, Q., et al. (2024). Learning adaptive shift and task decoupling for discriminative one-step person search. *Knowledge-Based Systems, 304*, Article 112483. https://doi.org/10.1016/j.knosys.2024.112483

Zhang, Q., Wu, J., Miao, D., et al. (2024). Attentive multi-granularity perception network for person search. *Information Sciences, 681*, Article 121191. https://doi.org/10.1016/j.ins.2024.121191

Zhang, Q., Miao, D., Zhang, Q., et al. (2025). Dynamic frequency selection and spatial interaction fusion for robust person search. *Information Fusion, 124*, Article 103314. https://doi.org/10.1016/j.inffus.2025.103314

Zhang, Y., Chen, M., Shen, J., et al. (2022). Tailor versatile multi-modal learning for multi-label emotion recognition. In: Proceedings of the AAAI Conference on Artificial Intelligence, pp 9100–9108, https://doi.org/10.1609/aaai.v36i8.20895

Zhao, Z., Liu, Q., & Wang, S. (2021). Learning deep global multi-scale and local attention features for facial expression recognition in the wild. *IEEE Transactions on Image Processing, 30*, 6544–6556. https://doi.org/10.1109/TIP.2021.3093397