scientific reports

Check for updates

OPEN

A text guided multimodal scale path fusion network for multimodal sentiment analysis

Siyuan Liu, Hongkun Zhao, Yang Chen, Fanmin Kong & Kang Li✉

Existing multimodal sentiment analysis (MSA) methods usually adopt fixed convolution kernels or static windows to model features from limited or fixed scales, making it difficult to dynamically model emotional features under different scale combinations. Furthermore, the absence of mechanisms to suppress redundant information in non-linguistic (video and audio) modalities hinders further performance improvements. To address these limitations, we propose a text guided multimodal scale path fusion network (TMSPF-Net). TMSPF-Net contains three main modules: Multi-scale Adaptive Transformer (MAT), Text-guided Conflict Elimination Module (TGCEM), and Channel Fusion Module. MAT captures the interaction of intra-modal and inter-modal through the combination of patches of different sizes and the dual attention mechanism, fully extracting multi-level global and local emotional information. Meanwhile, the adaptive routing module in MAT dynamically optimizes the feature paths through a learnable mechanism, enabling MAT to adaptively select the optimal path and increasing the flexibility of the model when dealing with heterogeneous data. TGCEM leverages multi-scale text-guided dynamic memory in MAT to filter conflicting signals and selectively preserve emotionally salient patterns in non-linguistic modalities, thereby improving the consistency and semantic richness of multimodal representations. Channel Fusion Module fuses the output results of these two modules and inputs them into the pre-trained language model to complete the MSA task. Extensive experiments on the MOSI and MOSEI datasets demonstrate that TMSPF-Net outperforms in most metrics than state-of-the-art methods. The results show that TMSPF-Net effectively guides the learning of non-linguistic modalities, integrates multi-level sentiment features, showing great potential in sentiment analysis.

In recent years, the research field of sentiment analysis has changed from single-modal to multimodal. The multimodal sentiment analysis (MSA) aims to integrate complementary data sources, including text, facial expression mode (vision), vocal music rhythm (audio), and other natural human communication channels[1], overall interpretation and analysis of emotional state. These heterogeneous modalities synergistically enhance model capability to capture nuanced affective states through distinct yet complementary cues[2]. For instance, textual data conveys explicit emotional semantics via lexical and syntactic structures, while facial microexpressions reveal implicit emotional fluctuations. Similarly, acoustic features such as pitch, intonation, and speech rate encode speaker-specific emotional tendencies, whereas phonetic characteristics like stress patterns and pauses provide additional insights into latent emotional dynamics[3,4].

Current research demonstrates that textual modalities exhibit superior contributions to multimodal sentiment analysis (MSA) compared to modalities[5–7]. Therefore, introducing pre-trained language model (PLM) into the model is an effective way to complete the MSA task[8,9]. Wang et al.[8] integrates multimodal outputs with PLM to enhance recognition accuracy but overlooks redundant information inherent in non-linguistic modalities. Subsequent studies address this limitation by augmenting CENet with a cross-modal redundancy elimination module and text-guided enhancement module[9]. However, the current work could be improved in two ways. First, the existing methods mainly extract the sentiment features of cross-modal interactions through multi-head attention mechanism, which can model the long-distance relationship between modalities. However, this approach often ignores the temporal granularity differences and multi-level emotional information in cross-modal dynamics. Second, while MCFNet improves representations through self-attention-based redundant suppression, it does not adequately address the persistence of intermodal conflicts in data (e.g., inconsistent audiovisual cues), which likewise degrades the model's performance.

School of Information Science and Engineering, Shandong University, Qingdao, Shandong, China. ✉email: kangli@sdu.edu.cn

Since the emotional information contained in different modalities of multimodal inputs at different scales may vary, multi-scale modeling is highly suitable for feature extraction or interaction as well as extracting multi-level emotional information[10–13]. Zhong et al.[10] and Yuan et al.[11] both adopted the idea of a fixed-scale sliding window to design sentiment analysis models in the task of dialogue sentiment analysis. Yue et al.[12] used convolution kernels of different fixed window sizes to extract the local features of the text respectively. Convolution kernels of different sizes can capture the features of n-gram text within different ranges. Wu et al.[13] proved for the first time that multi-scale feature fusion can effectively enhance emotional representation. Output proportionally sized features at three levels in a fully convolutional manner, and then aggregate the features of each level to construct an overall feature that is more related to emotions. Most multi-scale modeling methods adopt fixed convolution kernels or static window strategies to model features of limited or fixed scales, ignoring cross-scale interactions and multi-level emotional cues under different scale combinations. Meanwhile, different modal interaction features may prefer different scale combinations. Simply piling up scales may increase the model complexity and increase the model processing speed. Meanwhile, the fixed scale combination makes the interaction features can only be segmented according to the fixed scale, which leads to the poor generalization ability of the model for different types of data. Manually adjusting the optimal scale of each mode may not only fail to achieve the best performance but also be time-consuming and laborious. Therefore, designing models that can dynamically select different scales based on interaction characteristics is also our research focus.

To address these challenges, we propose a text guided multimodal scale path fusion network (TMSPF-Net). First of all, in terms of multi-scale modal feature interaction, we design a Multi-scale Adaptive Transformer (MAT), which can extract heterogeneous sentiment information between different modalities in more detail through patches of different scales combination. MAT set a series of different sizes of segmentation scale, different modalities can choose different scales of segmentation combination. The segmentation and combination of different scales can effectively extract more detailed emotional information between modalities. Subsequently, MAT employs a dual attention mechanism: intra-block attention to maintain local consistency within a single mode (e.g., word-level text features), and inter-block attention to model global cross-modal dependencies (e.g., discourse-level audio visual alignment). This hierarchical design enables granular emotion representation by jointly capturing intra-modal and inter-modal interactions[14,15]. At the same time, we integrated an adaptive routing module in MAT and utilized a learnable gating mechanism to autonomously optimize the fusion path during the training process. This dynamic architecture enhances adaptability to heterogeneous data structures[16,17]. In terms of redundancy elimination, we propose a Text-guided conflict elimination module which uses text embedding as a reference to reduce redundant or contradictory sentiment features in signals (audio/visual). By using different scales of text information to calculate the similarity matrix between text and non-linguistic pattern, if it is close enough to text information, it will be retained, and if it is not similar to text information, it will be discarded, so as to reduce the conflict between non-linguistic pattern and text information. In essence, it is to determine the dominant position of text in the model and guide the learning of non-linguistic modes through text information. Finally, the refined multi-scale representations are aggregated and processed by a pre-trained language model (PLM) to make the final sentiment prediction. The main contributions of this work are as follows:

- We designed a text guided multimodal scale path fusion network (TMSPF-Net), which enhanced the dynamic multi-scale cross-modal interaction and redundancy elimination.
- We designed a Multi-scale Adaptive Transformer (MAT) to capture local and global emotion information of different modalities by combining different scales and the dual attention mechanism. Meanwhile, MAT innovatively introduced Mixture of Experts (MoE) with the cross-modal attention mechanism, enhancing the modal interaction ability while also improving the flexibility of model path selection.
- We designed a Text-guided conflict elimination module (TGCEM). TGCEM leverages multi-scale text-guided dynamic memory to filter conflicting signals, and then eliminates the redundant and conflicting information in the non-linguistic modalities.
- We conducted a large number of experiments on two multimodal sentiment datasets. Our performance on the MOSEI dataset is better than the state-of-the-art (SOTA) model. On the MOSEI dataset, the regression metrics were improved by 1.0% and 1.1% compared with the SOTA method, and the seven-classification performance was improved by 0.88% and achieved an excellent performance of 55.68%.

Related work

In this section, section 2.1 introduces the relevant work of multi-scale feature extraction, section 2.2 introduces the relevant work of Transformer-based model, section 2.3 introduces the relevant work of Transformer-based pre-training model (PLM) in the field of multimodal sentiment analysis.

Multi-scale feature extraction

Traditional multimodal fusion approaches predominantly rely on feature concatenation or simplistic attention mechanisms, often neglecting critical disparities across modalities in temporal granularity (e.g., millisecond-level audio vs. word-level text sequences) and semantic hierarchy (e.g., lexical vs. discourse-level meaning). Recent advancements address these limitations through multi-scale feature extraction and adaptive fusion strategies.

As a cornerstone technique in deep learning[18–20], multi-scale feature extraction employs hierarchical architectures to capture temporal dynamics in multimodal signals. A seminal implementation by Lin et al.[21] demonstrates this through a top-down pyramidal structure that progressively constructs scale-specific feature representations, achieving state-of-the-art target detection accuracy through coordinated multi-resolution analysis.

Recent advances have seen growing adoption of multi-scale feature extraction in sentiment analysis. Zadeh et al.[14] pioneered this direction with their Multiple Attention Recursive Network (MARN), employing hierarchical attention mechanisms to model cross-modal interactions at both word and sentence levels, thereby demonstrating the efficacy of fine-grained temporal modeling for emotion classification. Addressing the challenge of extracting precise emotional cues from heterogeneous data while accounting for temporal dependencies between modalities, Gu et al.[22] developed a layered architecture integrating attention-level and word-level fusion for discourse-level emotion classification in text-audio pairs. Lei et al.[23] advanced speech synthesis through their MsEmoTTS framework, utilizing hierarchical modules to capture emotional prosody at varying temporal resolutions. Concurrently, Cao et al.[24] implemented parallel convolutional pathways with varying receptive fields for multi-label sentiment analysis, while Lin et al.[25] introduced channel-aware attention to model modality hierarchies, enabling dynamic cross-modal fusion. Building on these foundations, Fu et al.[26] extended the paradigm through multi-kernel convolutions and channel attention mechanisms, strategically prioritizing text modality features without compromising multimodal integration.

To optimize multi-scale feature fusion, contemporary approaches employ dynamic routing mechanisms that supersede static architectures with handcrafted connections. Sabour et al. pioneered dynamic routing mechanisms in Capsule Networks[27], enabling context-aware weight allocation between features through iterative agreement protocols rather than fixed topological constraints. Hazarika et al. advanced this paradigm in their MISA framework[28], decoupling modality-invariant and modality-specific representations while enforcing cross-granular semantic consistency through multi-scale contrastive learning objectives. Complementing these architectural innovations, Han et al.[17] pointed out in a review of dynamic neural networks that adaptive path selection based on gating can significantly improve the model's generalization ability to heterogeneous data.

Transformer-based model

Transformer is a sequence-to-sequence model originally applied to neural machine translation tasks that can efficiently model the transfer of information between long sequences. Transformer consists of two components, encoder and decoder, corresponding to the processing of source and target sequences respectively. Both the encoder and the decoder have a core component, Self-Attention, which enables the words in the sequence to obtain contextual information.

Paper[29] focuses on the problem that previous methods for fusing single-modal features into multimodal embeddings may lead to information loss or redundancy. A single mode enhancement transformer is introduced to extract the single mode information from the multi-mode embedding step by step and highlight the discriminant information. Paper[30] proposes a multimodal Transformer network based on rough set theory for emotion analysis and emotion recognition. This method improves the feature extraction and fusion ability of multimodal information through rough set self-attention and rough set cross-attention mechanisms. Paper[31] introduces a unified multimodal framework named UniMF, which aims to solve the problem of missing modes and alignment in multimodal sentiment analysis. The UniMF framework is particularly focused on dealing with multimodal sequences that are incomplete (that is, missing certain modes) and unaligned (that is, time or spatial synchronization issues between different modes), which are common challenges in real-world multimodal data collection. Paper[32] proposes a framework called TriSAT, which focuses on three-modal representation learning for multimodal sentiment analysis. The TriSAT framework is built on top of Transformer and introduces a module called Trimodal Multi-Head Attention (TMHA). This module treats language as the primary mode and combines information from visual and audio modes to enhance the accuracy of sentiment analysis.

Transformer-based pre-trained language model

The contribution of text modality in the sentiment analysis task is much greater than that of non-textual modalities. Therefore, the application of pre-trained language model in the field of multimodal sentiment analysis is increasing[33,34].

BERT is the most widely used pre-trained language model in the field of multimodal sentiment analysis. BERT's Transformer-based bidirectional encoder is pre-trained through mask language modeling and next sentence prediction. Its feature is to support context bidirectional understanding, which is suitable for natural language understanding tasks[35]. In 2019, RoBERTa (Robustly Optimized BERT Approach) was proposed as an improved version of BERT, which removes NSP tasks and optimizes training strategies (larger batch size, longer sequences), surpassing the original BERT on multiple NLU tasks with higher training efficiency[36]. SentiLARE (Sentiment-Aware Language Representation with Linguistic Knowledge) is a pre-trained language model for emotional perception proposed by Yequan Wang et al in 2020. Its core innovation lies in the explicit integration of linguistic knowledge (such as part of speech tagging, semantic role tagging) and affective information. By designing pre-training tasks for affective perception (such as affective word mask and sentence affective contrast learning), the model can enhance the modeling ability of textual affective polarity and fine-grained affective expression. It is significantly superior to traditional pre-trained models (such as BERT and RoBERTa) in tasks such as emotion classification and aspect level emotion analysis, and shows stronger robustness and generalization especially in low-resource scenarios[37].

Framework

The detailed architecture of TMSPF-Net is illustrated in Fig. 1. The modeling process is divided into three main components. First, multi-scale interactive emotional information across different modalities is extracted using the Multi-scale Adaptive Transformer (MAT). Second, redundant information among non-linguistic modalities is removed to enhance recognition accuracy. Finally, the outputs of the MAT and TGCEM are fused and fed into the PLM to perform the final sentiment analysis. In the following sections, we provide a detailed explanation of the structure and functionality of each module.

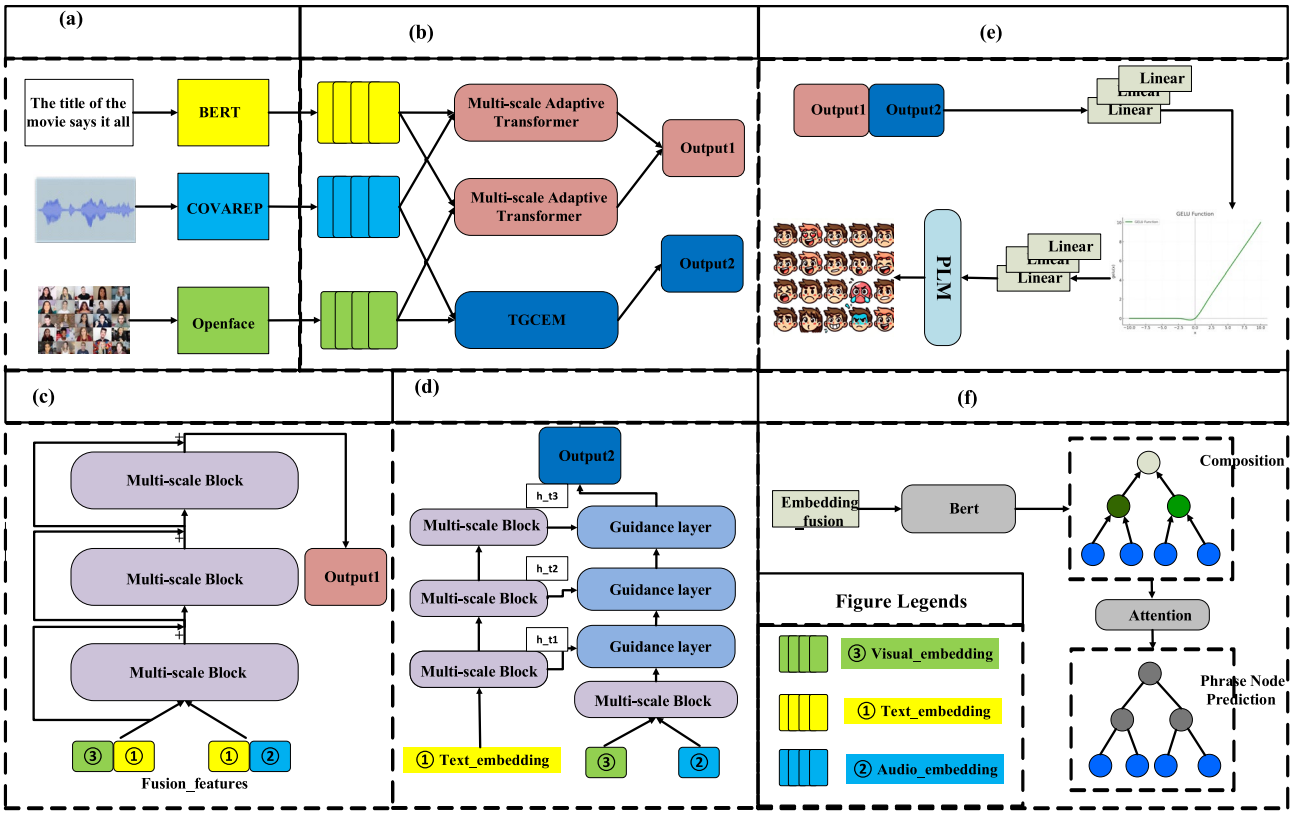


Fig. 1. The architecture of TMSPF-Net. (**a**) Input and feature extraction module. (**b**) Modalities interaction module. (**c**) Multi-scale adaptive transformer module. (**d**) Text-guided conflict elimination module. (**e**) Fusion layer. (**f**) Sentiment PLM produces the final sentiment prediction.

Multi-scale adaptive transformer

For the text, video and audio modalities of the initial input, we first use BERT, COVAREP and OpenFace to obtain the corresponding modal sentiment features f_t, f_a, f_v respectively, where $f_m \in [h_m, t_m, w_m]$ and m $\in$ {t,a,v}. Here, h_m denotes the feature channel dimension of modality m, t_m represents its temporal sequence length, and w_m corresponds to the auxiliary structural dimension of the modality. Since the features initially captured by BERT[35], COVAREP[38] and OpenFace[39] are not on the same dimension, it is not conducive to the interaction among features. Therefore, we project all the features into a common dimensional space and connect them pairwise based on text modalities, which will help obtain higher-quality information.

To more effectively extract and fuse multi-scale interaction information across different modalities, we propose a novel Multi-scale Adaptive Transformer (MAT), as illustrated in Fig. 1c. The goal of MAT is to dynamically extract multi-scale features from text-dominated mixed modalities. At the core of MAT is the Multi-scale Block (MSB), which plays a central role in capturing cross-modal dependencies. To address multi-scale dependency modeling in multimodal interactions, MSB incorporates a hierarchical adaptive feature fusion mechanism, as depicted in Fig. 2.

We define a set of patch sizes P={P_1, P_2,..., P_M}, where each P_i corresponds to a segmentation operation at a specific scale. For a given input feature X, each segmentation operation divides X into S patches, where $s = t/M$, resulting in a sequence {X_1, X_2,..., X_S}. Different patch sizes yield varying segmentation granularities, providing multi-resolution perspectives on the features. This completes the multi-scale segmentation stage.

Efficiently integrating these segmented features presents a key challenge. To address this, we propose a dual attention module that enables both inter-modal and intra-modal attention mechanisms. This design facilitates effective cross-modal feature interaction while preserving modality-specific information.

Intra-modal attention captures short-range dependencies by facilitating interactions between fine-grained features within local fragments. The core goal of intra-modal attention is to locally model the fine-grained features inside each patch, capture the context in the slice through the attention mechanism, and improve the precision of feature expression. The input of intra-modal attention is the previously divided S feature slices, such as {X_1, X_2,... X_S}, each feature slice contains several fine-grained emotional information. Taking the i-th patch $X_i \in [h, s, w]$ as an example, where h denotes the feature channel dimension, s represents temporal sequence length after being split by M patches, and w corresponds to the auxiliary structural dimension of the modality, we first embed along the feature dimension w, resulting in $x^i_{intra} \in [h, s, w]$. x^i_{intra} is then initialized using the query vector $Q \in [h, 1, w]$, which is usually initialized with trainable parameters to extract representative sentiment information from the feature film. We then design three linear layers through which we can map

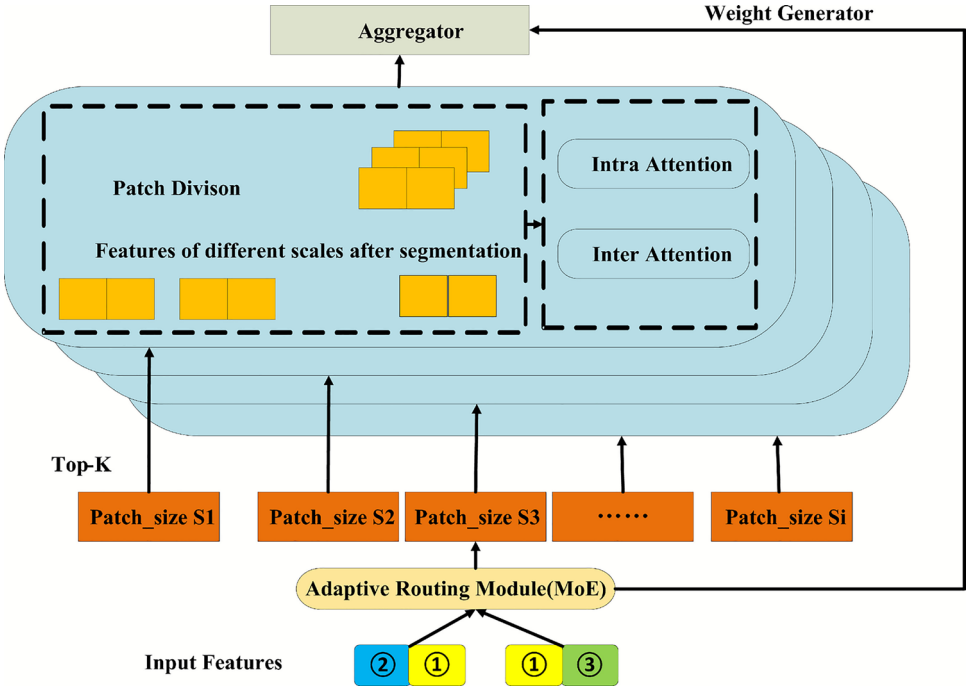


Fig. 2. Multi-scale block.

emotion features in feature slices into K^i_{intra}, V^i_{intra} and Q^i_{intra} vectors. Attention calculation for K^i_{intra}, V^i_{intra} and Q^i_{intra} vector features:

$$\text{atten}^i_{intra} = softmax(\frac{Q^i_{intra}K^i_{intra}}{\sqrt{D/H}})V^i_{intra}. \quad (1)$$

Softmax is a common normalization function that is mainly used to convert a set of arbitrary real numbers into a probability distribution. Each resulting $atten^i_{intra}$ is then concatenated to capture fine-grained time dependencies. Local context dependencies (short-range dependencies) within the capture feature slice can be captured. After all feature slices are given intra-modal attention, the output representation is Concat to form a fused representation for subsequent processing.

$$\text{atten}_{intra} = W_2^{shared} \cdot RELU(W_1^{shared} \cdot Concat * (atten^1_{intra} + atten^2_{intra} + \cdots + atten^i_{intra}) + b_1) + b_2. \quad (2)$$

ReLU is a widely used activation function, whose core idea is to enhance the expressibility of the model by nonlinear transformation of the input. Both W_1 and b_1 are trainable tensors.

The inter-modal attention module focuses on global dependency modeling between different feature slices, that is, capturing feature interactions over long distances and across modes. Compared with intra-modal attention, it focuses more on global modeling to improve the model's ability to understand complex emotional expression. The input features of inter-modal attention are similar to those of intra-modal attention. By flatting the mixed modal features in each feature slice into a unified vector representation, different feature slices can be uniformly input into the Transformer architecture for subsequent processing. Each flattened patch is translated to Query (Q_{inter}), Key (K_{inter}), and Value (V_{inter}) by three separate Linear layers. Then use Self-Attention to slice all the features into a common space to establish global dependencies:

$$\text{atten}_{inter} = softmax(\frac{Q_{inter}K_{inter}}{\sqrt{d_k}})V_{inter}. \quad (3)$$

This step can realize global information modeling across modal features, enhance the interaction between different feature slices, and improve the expression ability of emotional features. We then concatenate $atten_{intra}$ and $atten_{inter}$ to obtain the final attention output.

Meanwhile, we recognize that different scale partitions are suitable for different modality combinations. However, indiscriminately applying multiple scales may introduce redundant or irrelevant signals. To address this, we propose an Adaptive Routing Module that enables dynamic multimodal routing and multi-scale feature fusion. This module integrates features from various modalities and scales through a learnable gating mechanism. Notably, it combines the Mixture of Experts (MoE) architecture with cross-modal attention to enhance fusion flexibility. MoE improves model performance by employing multiple sub-networks ("experts"),

each of which specializes in handling different input patterns, thereby achieving more effective and adaptive information processing.

Each expert network focuses on learning different sub-regions of the input space, and the gated network acts as a "traffic controller" to achieve dynamic routing through the following equation:

$$y = \sum_{i=1}^{n} G_i(x) \cdot E_i(x). \tag{4}$$

Here, $E_i(x)$ represents the i-th expert network, and $Gi(x)$ represents the gating weight (which is usually satisfied with $\sum_{i=1}^{n} G_i(x) = 1$).

In order to avoid consistently selecting several patch sizes, resulting in the corresponding scales being repeatedly updated, while ignoring other potentially useful scales in the multiscale converter, we introduce a Top-K sparse gating mechanism for Gaussian noise:

$$\widetilde{g_i} = soft\max(\frac{W_g x + \varepsilon \cdot W_n}{K}), \tag{5}$$

where $\varepsilon \in$ N(0,1), K represents the first non-zero K values. At the same time, to prevent expert load imbalance, define variance loss:

$$L_{balance} = \lambda(CV(sum(gates)) + CV(load)). \tag{6}$$

CV is a calculation of Coefficient of Variation Squared. In the MoE, it is used to balance the load of different experts. Finally, multi-scale experts are used for parallel processing. Each expert is a Transformer Layer, whose window size is defined by patch size, and features of different scales are processed separately. The aggregator first performs transformation functions to align time dimensions from different scales. Then, the aggregator weights the multiscale output according to the path weights to obtain the final output:

$$MSB_{Out} = MLP(\sum_{i=1}^{K} g_i E_i(X) + X). \tag{7}$$

Text-guided conflict elimination module

In the MSA task, a significant challenge is the elimination of irrelevant or conflicting information from the visual and audio modalities. Such information refers to elements that are either unrelated to emotional expression or that contradict signals from other modalities. For example, in video data, background noise or extraneous visual effects may interfere with accurate emotion recognition. Similarly, in audio data, features that do not align with the speaker's true emotional state can lead to incorrect interpretations. To address this issue, we propose a Text-Guided Conflict Elimination Module (TGCEM), which leverages the semantic richness of textual information to guide the suppression of noise and inconsistencies in non-linguistic modalities.

The redundant information elimination module aims to remove emotional-irrelevant and conflicting information by guiding the learning of video and audio modalities using rich text information. This process helps enhance the model's performance and accuracy. The specific structure of this module is shown in Fig. 1d.

As can be seen from Fig. 1d, the text-guided conflict elimination module learns the features of text information at different scales independently through MAT, and guides the learning of non-linguistic features through text information at different scales.

We store the multi-scale text features extracted by MAT in an $h_t^i \in$ [h, t, w]. Then we build the non-linguistic modal guidance layer dynamically through text semantic guidance. In the initialization phase, we design an initial text guidance feature $h_{guidance}^0 \in$ [1, t, w] for text features, and broadcast it to the batch dimension [h, t, w]. For the *j*-th guidance layer, we define the input as the text features h_t^i, audio features h_a, video features h_v and parameters of the previous layer $h_{guidance}^{j-1}$. Then we calculate the similarity proof between text and non-linguistic modalities. The composition of the guidance layer is shown in Fig. 3.

We input the output of text embedding after MAT h_t^i as Query input, which we define as Q_{Te}^i. Similarly, audio feature and video feature are input into the first guidance layer as Key and Value after the same operation. Then, the similarity relationship between text mode and audio mode, text mode and video mode is calculated respectively to update the non-linguistic modal features. Formula 8 represents the similarity relationship between text and audio modes.

$$\alpha = softmax(\frac{Q_{Te}^i {K_{Au}}^T}{\sqrt{d_k}}). \tag{8}$$

The Softmax function is an activation function that normalizes a numerical vector to a probability distribution vector. N is the dimension of each attention head. Similarly, the similarity matrix between text and video modes is calculated as follows:

$$\beta = soft\max(\frac{Q_{Te}^i {K_{Vi}}^T}{\sqrt{d_k}}). \tag{9}$$

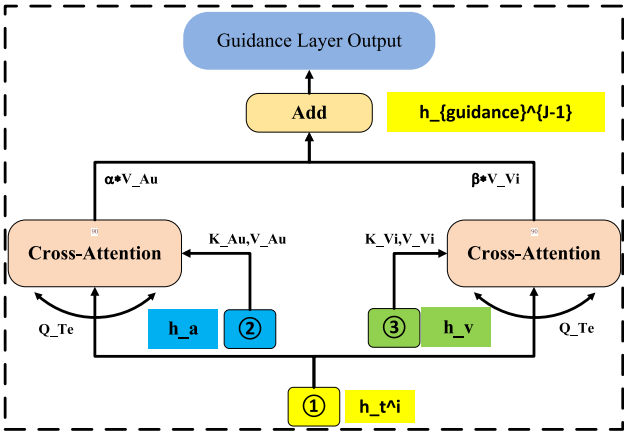


Fig. 3. The composition of the guidance layer.

In summary, the output of the guidance layer can be iterated by constantly updating the sum of the similarity matrix:

$$Output_2^j = h_{guidance}^{j-1} + \alpha V_{au} + \beta V_{vi}. \tag{10}$$

From the output of the guidance layer, we believe that the non-affective information in the mode has been filtered out, and the generated output is a mode information that contains rich emotional information and is sufficiently refined.

Channel fusion

Since the outputs of MAT and TGCEM are not at the same dimension, we unified the output shapes of MAT and TGCEM through a series of operations including linear layer, activation function and linear layer, and then combined the outputs of these two modules into the SentiLare pre-training model as training inputs. The composition of the fusion layer is shown in Fig. 1e. Since both of these modules are learned by text modality, pre-training language models can well analyze the emotional representations in them, and thus produce the final emotion analysis.

Experiment

Datasets

The experimental data was selected from the MOSI[40] dataset and MOSEI[41] dataset. MOSI is a multimodal sentiment analysis dataset released in 2016, which includes 2198 video clips, single shot commentary videos on YouTube, and text of each short video recorder's audio content. In 2018, researchers released the MOSEI dataset, a massive sentiment analysis dataset with 22856 video clips that was also obtained from YouTube. Higher values indicate stronger positive sentiment polarity. The sentiment scores for each video clip in the MOSI and MOSEI datasets fall within the [-3, 3] interval.

Evaluation metrics

MAE, Corr, ACC-2, F1 and ACC-7 were selected as evaluation indicators of the model. In the regression task, MAE and Corr are selected as evaluation metrics. MAE is the average absolute error, which represents the average absolute error between the predicted value and the true value. The smaller the value, the more accurate the model prediction. Corr is the Pearson correlation coefficient between the predicted value and the true value, and its function is to measure linear correlation. In the classification task, ACC-2, F1 and ACC-7 are selected as evaluation metrics. ACC-2 represents the accuracy of binary classifications, and ACC-7 represents the accuracy of seven classifications. F1 is the harmonic average of Precision and Recall. Precision is the proportion of samples with positive prediction that are actually positive, and recall is the proportion of samples with positive prediction that are correctly predicted. The higher the value, the better the model will be.

Experimental details

The configuration of the experimental environment mainly refers to the previous experiment[8,9]. All of our experiments were conducted on the Tesla V100S PCIE32GB GPU. The Py-Torch model framework is adopted. The environment configuration is python v3.8 (https://www.python.org/downloads/release/python-380/) + pytorch v1.8.1 (https://pytorch.org) + cuda v11.1 (https://developer.nvidia.com/cuda-toolkit-archive), and the software used is PyCharm Community Edition v2024.2.1 (https://www.jetbrains.com/pycharm/). The optimizer is selected as Adam, and the Adam epsilon parameter is set to 1e-8, which is consistent with the Settings of most models. The number of epoch is set to 30. The learning rate is set at 6e-5. The main hyperparameters in TMSPFNet are shown in Table 1.

Parameters	CMU-MOSI	CMU-MOSEI
Batch size	64	32
Seq length	50	50
Number of layers in MAT	2	2
Number of layers in TGCEM	3	3
Dropout	0.5	0.5
K	4	4

Table 1. Experimental details.

Baseline

We conducted a comprehensive comparative study on TMSPF-Net, and our research baseline not only selected early typical models, but also compared them with different types of excellent models in recent years.

1. M3SA: Multi-Scale Feature Extraction and Multi-Task Learning (Multi-Scale method, 2024)[25].
2. TMFN: A text-centric multimodal fusion network that leverages multi-scale feature extraction, channel attention mechanisms, and unsupervised contrastive learning (Multi-Scale method, 2025)[26].
3. TFPN: A bidirectional text-guided model based on feedback gated progressive fusion and cross-temporal attention mechanism (Text-guided method, 2025)[42].
4. ALMT: Learning Language-guided Adaptive Hyper-modality Representation (Text-guided method, 2023)[43].
5. SmartRAN: A model based on the dynamic selection of data flow paths by the intelligent routing attention module and the joint optimization of the dual-flow learning mechanism inside and outside the modal (Intelligent routing method, 2024)[44].
6. TGMoE: A model based on a text-guided cross-modal attention mechanism and a sparse gated Mixture-of-Experts (Intelligent routing method, 2024)[45].
7. CENet: A model based on cross-modal enhancement modules and feature transformation strategies (Transformer-based PLM method, 2023)[8].
8. MCFNet: A model based on a multi-channel cross-modal fusion framework, combined with language information enhancement and auxiliary modal redundancy elimination (Transformer-based PLM method, 2024)[9].
9. MST-ARGCN: Transformer-based model with attentional recurrent graph capsule network[46], 2025.
10. MMAFN: A transformer-encoder-based model[47], 2024.
11. CMHFM: Cross modal hierarchical fusion based on multi-task learning[48], 2024.
12. TMBL: A multimodal framework based on Transformer[49], 2024.
13. PS-MIXER: First application of MLP-Mixer in tri modal sentiment analysis framework[50], 2023.
14. DHCN: A novel hypergraph neural network[51], 2024.
15. NUAN+: A non-uniform attention module with combined loss for multi-modal sentiment analysis that fuses text, audio, and visual features into a tripartite interaction representation[52], 2025.
16. NUAN: A text-centric non-uniform attention network that integrates acoustic and visual modalities into LSTM-based recurrent feature fusion for sentiment prediction[53], 2022.
17. MulT: A model based on multimodal transformer and directed pairwise cross-modal attention[11].
18. LMF: A model based on low-rank multimodal fusion[4].
19. MFN: A model based on memory fusion network[2].
20. TFN: Tensor Fusion Network[3].

Quantitative analysis

The experimental results of MSA tasks are shown in Table 2. On these two datasets, TMSPF-Net was compared with multiple baseline models and achieved SOTA effect on most metrics.

On the MOSI dataset, TMSPF-Net fully achieved the SOTA performance in the regression metrics. The MAE reached 0.584 and the Corr reached 0.874, both increasing by 1.5% compared with the suboptimal model CENet. In the classification task, only the binary classification was slightly lower than that of the suboptimal model MCFNet, but still can achieve an excellent accuracy rate of 90.02%. Meanwhile, both the F1 and ACC-7 indicators achieved the optimal performance, Among them, F1 increased by 0.39% and ACC-7 increased by 0.8%. This indicates that the model can effectively capture subtle sentiment changes. Meanwhile, in the many experiments we conducted, two superior performances emerged. Among them, MAE reached 0.567, Corr reached 0.881, ACC-2 reached 91.33%, F1 reached 91.22%, and the performance of ACC-7 reached 50.98%. Compared with the suboptimal model, MAE and Corr increased by 4.4% and 2.3% respectively, and the binary classification index increased by 1.09%. The seven-classification index has increased by 1.28%. At this point, the performance of the model has significantly improved in both regression tasks and classification tasks. However, the frequency of occurrence of this type of data is relatively low and it has a certain degree of uncertainty. Therefore, we did not show it in Table 2. However, the emergence of such data can still indicate the performance and potential of our model to a certain extent.

On the MOSEI dataset, TMSPF-Net achieved SOTA performance in both regression and classification metrics. In terms of MAE, TMSPF-Net achieved the highest 0.503, which was nearly 1% higher than the suboptimal model SmartRAN. In terms of Corr, TMSPF-Net achieved the highest score of 0.806, which was

Models	MOSI					MOSEI				
	MAE↓	Corr↑	ACC-2↑	F1↑	ACC-7↑	MAE↓	Corr↑	ACC-2↑	F1↑	ACC-7↑
LMF	0.912	0.688	76.4	75.7	32.8	0.623	0.677	82.0	82.2	48.0
TFN	0.970	0.633	73.9	73.4	32.1	0.593	0.700	82.5	82.5	50.2
MFN	0.965	0.632	77.4	77.3	34.1	0.568	0.717	84.4	84.3	51.3
MulT	0.871	0.698	83.0	82.8	40.0	0.630	0.664	80.1	80.9	49.0
PS-MIXER	0.794	0.748	82.1	82.1	44.3	0.537	0.765	86.1	86.1	53.0
TMBL	0.867	0.762	83.84	84.29	36.3	0.545	0.766	85.8	85.92	52.4
DHCN	0.899	0.699	82.5	82.7	38.6	0.6033	0.685	81.8	81.9	50.13
NUAN	1.034	0.654	78.3	77.9	26.8	0.624	0.653	81.2	80.6	48.6
NUAN+	0.979	0.649	78.7	78.6	30.9	0.744	0.624	79.7	78.4	42.5
CMHFM	0.912	0.677	81.0	81.3	37.0	0.560	0.733	84.09	83.83	52.39
ARGCN	0.925	0.659	81.3	81.5	36.0	0.598	0.681	81.8	82.2	50.5
MMAFN	0.830	0.740	84.30	84.19	43.30	0.554	0.748	85.50	85.26	53.25
TMFN	0.709	0.791	85.28	85.37	/	0.531	0.742	86.10	86.17	/
M3SA	0.7133	0.801	84.72	86.61	/	/	/	/	/	/
TFPN	0.687	0.811	87.5	87.51	49.7	0.528	0.771	85.73	85.86	54.8
ALMT	0.683	0.805	86.43	86.47	49.42	0.526	0.779	86.79	86.86	54.28
SmartRAN	0.684	0.810	87.07	87.04	46.69	0.508	0.797	87.30	87.23	54.7
TGMoE	0.760	0.767	85.64	85.71	45.89	0.535	0.757	85.51	85.86	53.70
CENet*	0.593	0.861	89.78	89.24	49.5*	0.541	0.792	87.15	86.23	54.2*
MCFNet*	0.601	0.859	**90.24**	89.62	49.6*	0.537	0.796	87.23	86.82	54.8*
TMSPF-Net	**0.584**	**0.874**	90.02	**90.01**	**50.5**	**0.503**	**0.806**	**87.67**	**87.37**	**55.68**

Table 2. Experimental results of TMSPF-Net and baseline models on MOSI and MOSEI datasets. The best results are indicated in bold. Data with * are reproduced by us.

1.1% higher than the suboptimal model. In the classification task, ACC-2 and ACC-7 can reach 87.67% and 55.68% respectively, which are 0.37% and 0.88% higher than the suboptimal model respectively. To sum up, TMSPF-Net achieved SOTA performance in nine out of the ten evaluation metrics of the two multimodal sentiment analysis datasets.

Compared with typical baseline models such as TFN and MULT, the performance of our proposed model on both MOSI and MOSEI data sets has been greatly improved, which is caused by the following reasons: 1. The simple concatenation of multimodal features does not take into account the main contribution of text modality. 2. typical baseline models do not consider the different sentiment information in the features at different scales, and it is unable to capture subtle emotions, which also leads to the low of Acc-7. 3. Less use of pre-trained language models.

Compared with M3SA, TMFN (multi-scale method), TMSPF-Net also improved in all metrics. On the MOSI dataset, compared with M3SA, TMSPF-Net improved by 5.3% in the binary classification and by 3.4% in the F1 metric. Compared with TMFN, TMSPF-Net improved by 4.74% in the binary classification and by 4.64% in the F1. In terms of the MAE, TMSPF-Net decreased by 17.6% compared with TMFN, which fully indicates that the accuracy of TMSPF-Net is higher. On the MOSEI dataset, TMSP-FNet also demonstrated excellent performance. We believe that the pre-trained language model obtains an enhanced representation of text information, thereby improving the accuracy of sentiment analysis tasks, and that the text-led learning mode is also an important factor in improving model performance. Meanwhile, compared with the fixed convolution kernels and fixed window scale division adopted by TMFN and M3SA, the multi-scale combination method of MAT has demonstrated excellent performance in coarse-grained classification.

Compared with text-guided methods, such as ALMT and TFPN, TMSPF-Net not only has a higher improvement in the binary classification, but also has made progress in the seven-classification. Compared with ALMT, TMSPF-Net improved by 1.08% in the seven-classification. In terms of the MAE, TMSPF-Net decreased by 14.4%. In terms of the Corr, TMSPF-Net improved by 8.5%, which indicates that the model at this time is both accurate and captures the trend. On the MOSI dataset, compared with TFPN, TMSPF-Net improved by 2.52% in the binary classification and by 0.8% in the seven-classification. Compared with TFPN on the MOSEI dataset, TMSPF-Net improved by 1.94% in the binary classification and by 0.88% in the seven-classification. This indicates that multi-scale feature extraction can effectively capture subtle emotional features and effectively promote the improvement of seven classification indicators. TMSPF-Net is also superior to intelligent routing methods such as SmartRAN and TGMoE.

Compared with the pre-trained language models, TMSPF-Net also achieved the SOTA effect in most metrics. Whether on the MOSI or MOSEI datasets, TMSPFNet outperforms MCFNet in both MAE and Corr metrics, and has a nearly 1% performance improvement in the ACC-7 metric. This indicates that TMSPFNet can effectively capture fine-grained sentiment changes and improve the quasi-determinism of multiple classifications. This is mainly due to the multi-scale transformer we proposed.

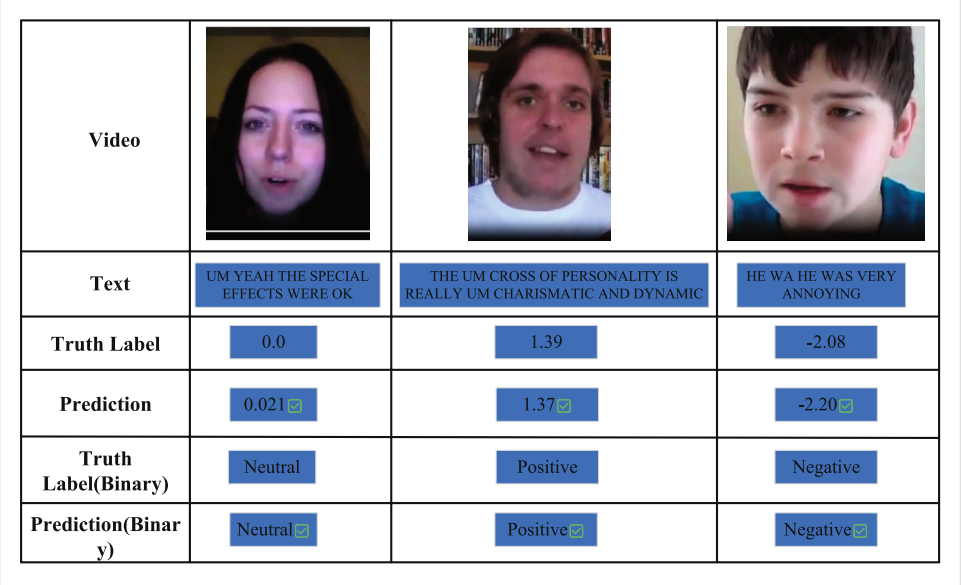


Fig. 4. Three groups of fragments were selected for case analysis in the MOSI test set.

Modalities	MOSI					MOSEI				
	MAE↓	Corr↑	ACC-2↑	F1↑	ACC-7↑	MAE↓	Corr↑	ACC-2↑	F1↑	ACC-7↑
V	1.503	0.196	57.7	25.0	15.5	0.835	0.227	63.5	60.5	40.0
T	0.785	0.789	83.8	83.7	41.9	0.618	0.735	85.5	85.5	53.7
A	1.528	0.149	57.7	50.7	15.3	0.858	0.191	65.2	62.3	37.2
V+T	0.603	0.857	88.8	88.8	48.2	0.510	0.802	87.5	87.2	55.2
V+A	1.437	0.127	61.7	60.6	18.6	0.859	0.239	66.1	60.9	38.1
A+T	0.598	0.862	89.3	89.3	48.1	0.517	0.796	86.9	86.5	54.4
V+A+T	**0.584**	**0.874**	**90.02**	**90.01**	**50.5**	**0.503**	**0.804**	**87.67**	**87.37**	**55.68**

Table 3. Influence of different modality combinations on MOSI and MOSEI datasets. The best results are indicated in bold.

Case study

To verify the predictive performance of our model, we conducted a multimodal case analysis using three video clips from the test set of the CMU-MOSI dataset, as shown in Fig. 4. Each case presents aligned multimodal features, namely text and video clips, basic truth labels, model predicted values [-3, 3], and the corresponding binary classification model predicted values.

Case 1 demonstrates accurate neutral sentiment detection. Textual input: "THE SPECIAL EFFECTS WERE OK" (neutral), Prediction output: 0.021 (Truth label: 0.0), Binary classification: Neutral. Case 2 shows precise positive sentiment capture. Textual input: "THE CROSS OF PERSONALITY IS REALLY CHARISMATIC AND DYNAMIC" (Positive), Prediction: 1.37 (Truth label: 1.39), Binary classification: Positive. Case 3 reveals negative sentiment recognition. Textual input: "HE WAS VERY ANNOYING" (Negative), Prediction: -2.20 (Truth label: -2.08), Binary classification: Negative.

Based on the above cases, it can be found that the difference between the predicted value and the Truth label is very small, which can well achieve sentiment analysis in different states. This also proves the effectiveness of TMSPF-Net.

Ablation experiment

Influence of different modalities combinations

We first verify the effect of the combination of different modes on the MSA task of the model. We conducted single-mode (text, video, audio), dual-modal (text + video, text + audio, video + audio) and tri-modal (text + video + audio) experiments on MOSI and MOSEI datasets, respectively. The experimental results are shown in Table 3.

It can be seen from Table 3 that for a single pattern, the experimental results obtained by sentiment analysis only through text are the best. On the MOSI dataset, the binary classification accuracy rate can reach 83.8% only through text modality, and on the MOSEI dataset, the binary classification accuracy rate can reach 85.5% only through text modality. Therefore, compared with audio and video, text has the greatest impact on the results of

Module	MOSI					MOSEI				
	MAE↓	Corr↑	ACC-2↑	F1↑	ACC-7↑	MAE↓	Corr↑	ACC-2↑	F1↑	ACC-7↑
ONLY MAT	0.589	0.863	88.5	88.5	50.5	0.507	0.793	86.5	86.5	55.4
ONLY TGCEM	0.590	0.864	87.6	87.6	48.7	0.510	0.790	85.87	85.76	55.0
MAT + TGCEM	**0.584**	**0.874**	**90.02**	**90.01**	**50.5**	**0.503**	**0.806**	**87.67**	**87.37**	**55.68**

Table 4. Influence of different module combinations on MOSI and MOSEI datasets. The best results are indicated in bold.

Models	MAE↓	Corr↑	ACC-2↑	F1↑	ACC-7↑
MCFNet-original*	0.583	0.864	89.1	89.1	49.6
MCFNet-MAT	0.578	0.872	90.0	90.0	50.0
MCFNet-TGCEM	0.583	0.867	89.3	89.2	50.2
CENet-original*	0.588	0.862	88.1	88.0	49.5
CENet-MAT	0.580	0.866	89.75	89.7	49.8
CENet-TGCEM	0.587	0.864	89.3	89.2	49.7

Table 5. The effectiveness of the MAT and TGCEM in various models.

sentiment analysis. For dual-modal binding, it can be seen that the experimental results obtained by text-based modal binding are much higher than those of non-text-based interaction methods, which further proves the importance of text features in the task of sentiment analysis. Compared with the modal effect of binding text and audio, the model of binding text and video has a higher accuracy rate, which indicates that the emotional information in the video modality is richer compared with the speech modality. Compared with the single-modal binding method, the results of the dual-modal binding method have also been improved, which also proves that there is complementary information between the modalities. The mutual combination of patterns can improve the accuracy of sentiment analysis tasks. The experimental results obtained from the three-modal binding are the most superior, which also proves the effectiveness of the three-modal binding of text + video + audio.

Influence of different module combinations

In this section, we verify the effect of different module combinations in the model on the model MSA task. We conducted single-module (MAT, TGCEM) and dual-module binding combination experiments on MOSI and MOSEI datasets respectively. The experimental results are shown in Table 4.

As can be seen from Table 4, For a single module, excellent performance can still be achieved with only MAT. On the MOSI dataset, using only the MAT module, the binary classification accuracy rate can reach 88.5%, and the seven-classification accuracy rate can reach 50.5%. It is indicated that multi-scale feature extraction can fully capture fine-grained and coarse-grained emotional features. Although the performance of the model with only TGCEM is inferior to that with only MAT, there is no phenomenon that sentiment analysis cannot be carried out when only nonverbal modes are used as shown in Tables 3, and an effective output can still be achieved. This shows that the TGCEM module makes the output of the non-linguistic modes close enough to the text mode and eliminates the suppression and conflict information in the non-linguistic modes. After the combination of the two modules, the optimal experimental results are obtained, which also proves that MAT and TGCEM can complement and cooperate with each other to obtain the best experimental results.

Qualitative analysis

This section explores the effectiveness of the MAT and TGCEM in various models.

The TGCEM is designed to reduce information redundancy during multimodal fusion. MAT uses multi-scale feature extraction to extract global and local subtle emotional features between mixed modes, and promotes the interaction between text modes and multimodal features. To test the mobility and robustness of these two modules, we integrate the MAT and TGCEM into the CENet and MCFNet models respectively, where MCFnet-original and CENet-original are indicators of MCFNet and CENet under the same experimental conditions.

It can be seen from Table 5 that the performances of both MCFNet and CENet have improved to varying degrees. On the MCFNet model, the MAT module was used to replace the original self-attention module, and the binary classification and seven-classification of the model increased by 0.9% and 0.4% respectively. Replacing the redundancy elimination module in MCFnet with the TGCEM module results in a relatively small improvement in the binary classification performance of the model, but a relatively high improvement in the seven-classification accuracy. On the CENet model, the MAT module and the TGCEM module have a more obvious improvement in the model performance. This indicates that the two modules can adapt to different model scenarios and have a greater impact on the accuracy of the seven-classification, proving that the model has good robustness.

The influence of number of MSB in MAT

In this section, we tested the influence of the number of MSB in MAT on the model performance. We conducted tests for the four situations when the number of MSB = 1, the number of MSB = 2, the number of MSB = 3, and the number of MSB = 4 respectively. The training time was newly added as a reference in the evaluation metric. The experimental results are shown in Table 6.

The MAE achieves the optimal performance of 0.584 when the number of MSB is 2, while the Corr achieves the optimal performance of 0.882 when the number of MSB is 3. When the number of MSB is 3, the binary classification can achieve an outstanding performance of 90.41%, but at this time, the 7-classification is 50.1%, slightly lower than the 50.5% of the 7-classification when the number of MSB is 2. The changes of the above metrics all declined when the number of MSB was 4, which indicates that simply stacking the number of MSB does not necessarily enable the model to achieve the optimal performance. Based on the changes of the comprehensive classification and regression metrics, the number of MSB can be set to 2 or 3, both of which can achieve outstanding performance higher than that of the SOTA method. However, we added the Training Time as an important evaluation. We can see that as the number of MSB increases, the Training Time multiplies. When the number of MSB is 3, The Training Time is 4519 seconds, which is much higher than the Training Time when the number of MSB is 2. To sum up, in order to balance the model effect and efficiency, we set the number of MSB to 2.

The influence of number of guidance layer in TGCEM

We also verified the impact of the number of Guidance Layer in TGCEM on the model performance. We selected different layers to conduct corresponding ablation experiments. The trends of each metric changing with the number of layers are shown in Fig. 5.

Figure 5a shows the changes of MAE and Corr with the number of Guidance layers. The range of the blue coordinate axis represents the variation range of MAE. The orange coordinate axis represents the variation range of Corr. As shown in Fig. 5a, when the number of Guidance layers is 3 and 4, the MAE is optimal and can reach 0.584. When the number of Guidance layers is 3, the Corr reaches a peak of 0.874. To sum up, when the number of Guidance layers is 3, the metrics of the regression task reaches the optimum. Figure 5b shows the changes of the classification metrics Acc-2 and F1 with the number of Guidance layers. It can be clearly seen from Fig. 5b that when the number of Guidance layers is 3, the binary classification and F1 performance are the best, reaching 90.02% and 90.01% respectively. When the number of Guidance layers is 4, the binary classification shows a decrease. Figure 5c shows the changes of Acc-7. It can be clearly seen from Fig. 5c that with the increase of the number of Guidance Layer layers, the accuracy rate of Acc-7 gradually improves. When the Guidance Layer = 4, it reaches the highest 51.16%. Based on the above experimental results, we choose to set the Guidance Layer to 3 to balance the comprehensive performance of the regression task and the classification task.

The influence of different scale decomposition

In this section, we verify the influence of different scale decomposition methods in the model on the experimental results. Since the sequence length is set to 50, the selected decomposition scale must be divisible by 50. Therefore, we set the decomposition scale as follows [50,25,10,5,2,1]. MAT adaptively selects the first K patch sizes to combine to adapt to different time series samples. We evaluated the influence of different K values on the prediction accuracy in Table 7. Our results show that the results of K = 2 and K = 3 are better than those of K = 1 and K = 4, highlighting the advantages of adaptive modeling of key multi-scale features to improve accuracy. When K = 5, the seven classification es and mae of the model are both improved, but the training time is also longer, so we choose k = 4 after comprehensive consideration. Different modalities benefit from feature extraction using different patch sizes, but not all patch sizes are equally effective.

Visualization of different expert-scale weights

We show in Fig. 6 the preferences of different channel features for different Expert-scales. As can be seen from Fig. 6a, the samples of channel 10 and channel 11 can respectively achieve weights of 0.42 and 0.49 for Expert 4, indicating that these two samples are more suitable for division using the segmentation scale of Expert 4. The weight of channel 9 reaches the highest on the scale of Expert 1, while the samples of channel 12 prefer Expert 3 more. Figure 6b shows that each channel sample has its preferred Expert scale. These observations emphasize the adaptability of MoE and its ability to identify the optimal combination of plaque sizes for different modalities.

The influence of learning rate

In this subsection, we tested the influence of different learning rates on the model performance. We choose the following learning rates for evaluation: 1e-5, 3e-5, 6e-5, 1e-4, and 6e-4. Table 8 shows the variation of model performance under different learning rates.

Nums of MSB	MAE↓	Corr↑	ACC-2↑	F1↑	ACC-7↑	Training time↑
MSB layers = 1	0.586	0.864	88.72	88.68	49.7	**497**
MSB layers = 2	**0.589**	0.874	90.02	90.01	**50.5**	**624**
MSB layers = 3	0.598	**0.882**	**90.41**	**90.31**	50.1	4519
MSB layers = 4	0.594	0.859	89.63	89.62	49.125	8770

Table 6. The influence of number of MSB in MAT. The best results are indicated in bold.

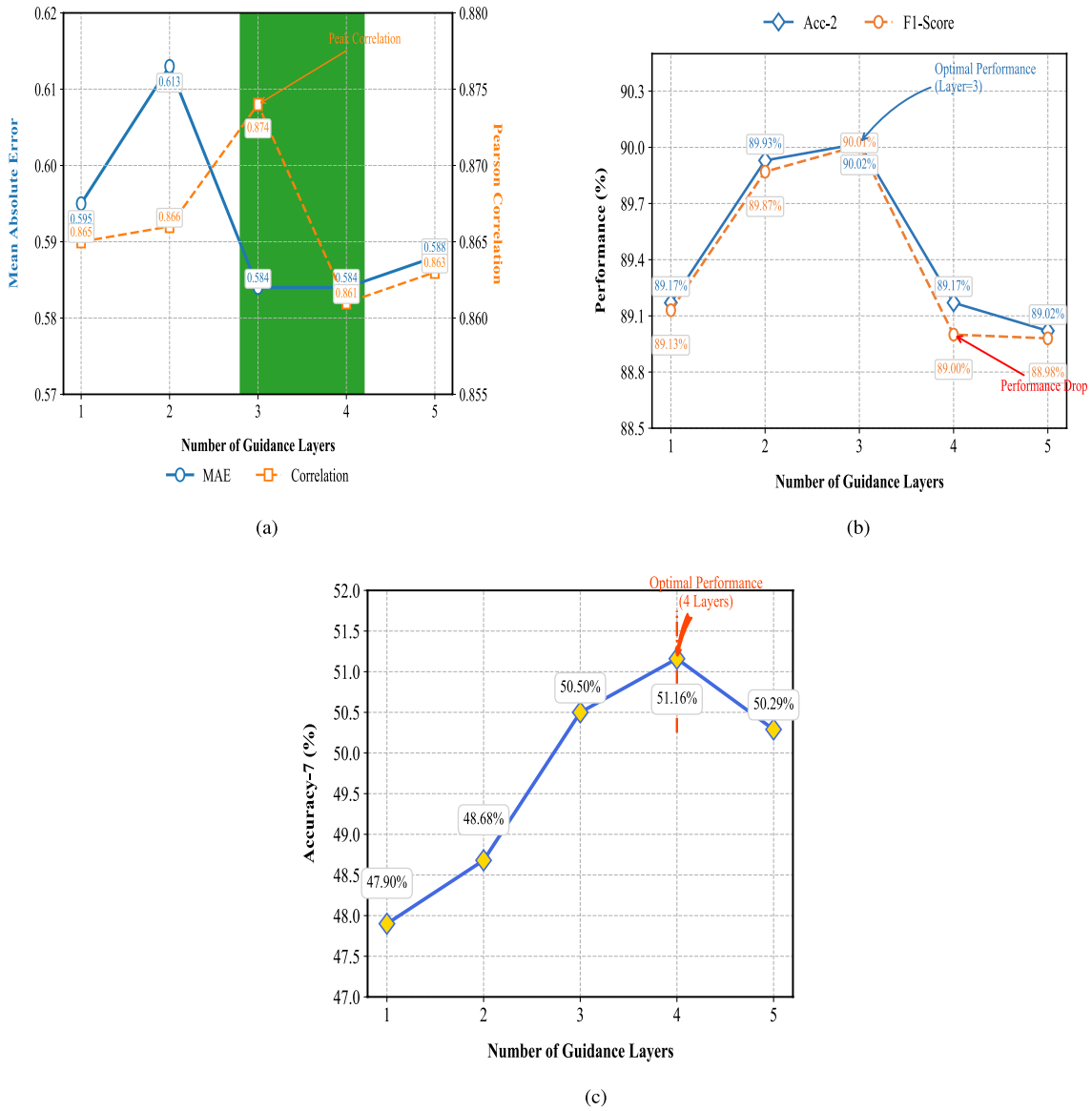


Fig. 5. The influence of the number of guidance layers in TGCEM.

Nums of K	MAE↓	Corr↑	ACC-2↑	F1↑	ACC-7↑
K =2	0.589	0.861	89.0	88.9	48.9
K =3	0.589	0.863	89.7	89.6	50.2
K =4	0.584	**0.874**	**90.02**	**90.01**	50.5
K =5	**0.579**	0.852	88.72	88.63	**51.02**
K =6	0.590	0.858	87.6	87.6	51.05

Table 7. The influence of different scale decomposition on MOSI dataset. The best results are indicated in bold.

It can be seen from Table 8 that the MAE first decreases as the learning rate increases and reaches the optimum when the learning rate is 6e-5. After that, MAE gradually increases as the learning rate becomes larger. For example, when the learning rate reaches 6e-4, the MAE rises to 0.776. The Corr metric exhibits a similar trend: it increases with the learning rate and reaches its maximum at 6e-5, and then decreases when the learning rate continues to grow. When the learning rate is 6e-4, Corr drops to 0.751. Therefore, in terms of regression performance metrics, the learning rate of 6e-5 provides the best overall results. For classification performance, the binary accuracy (ACC-2) and F1-score reach their highest values when the learning rate is 3e-5, which are

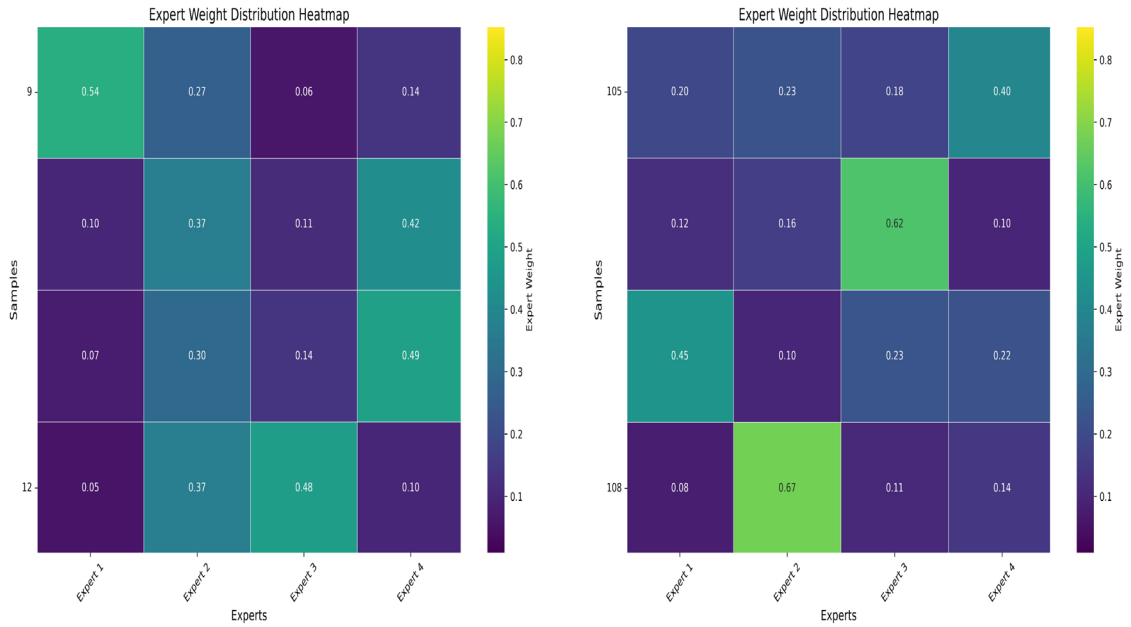


Fig. 6. Visualization of different expert-scale weights.

Different learning rate	MAE↓	Corr↑	ACC-2↑	F1↑	ACC-7↑
lr = 1e-5	0.607	0.856	88.41	88.39	50.43
lr = 3e-5	0.590	0.863	**90.24**	**90.20**	50.14
lr = 6e-5	**0.584**	**0.874**	90.02	90.01	**50.5**
lr = 1e-4	0.602	0.859	88.56	88.49	48.52
lr = 6e-4	0.776	0.751	84.14	83.91	42.42

Table 8. The influence of different learning rate. The best results are indicated in bold.

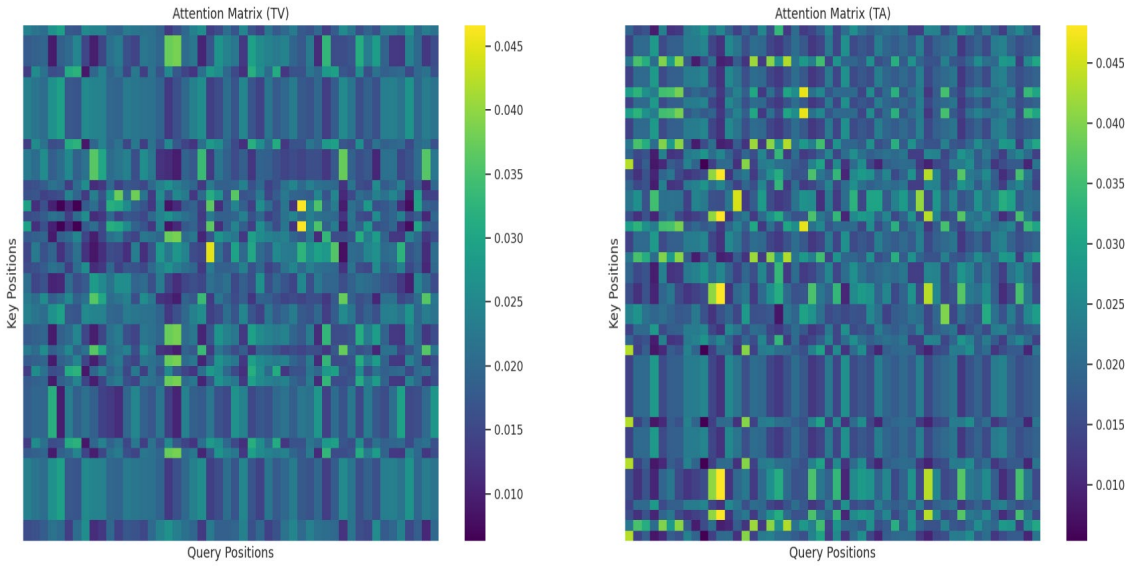
slightly higher than those obtained at 6e-5. The seven-class accuracy (ACC-7), however, achieves its maximum value of 50.5% at 6e-5. Another common phenomenon shown in Table 8 is that when the learning rate exceeds 1e-4, the model performance experiences significant degradation. We believe that an excessively large learning rate negatively affects the model's optimization process. To sum up, considering both regression and classification metrics, we choose 6e-5 as the optimal learning rate.

Average attention matrix

In Fig. 7, we give the average attention matrix (i.e., alpha and beta). As shown in Fig. 7, the redundant information elimination mode pays more attention to visual modality, indicating that visual modality provides more complementary information than auditory modality. In addition, as can be seen from Table 3, the model performance degrades more significantly when the video input is removed than when the audio input is removed.

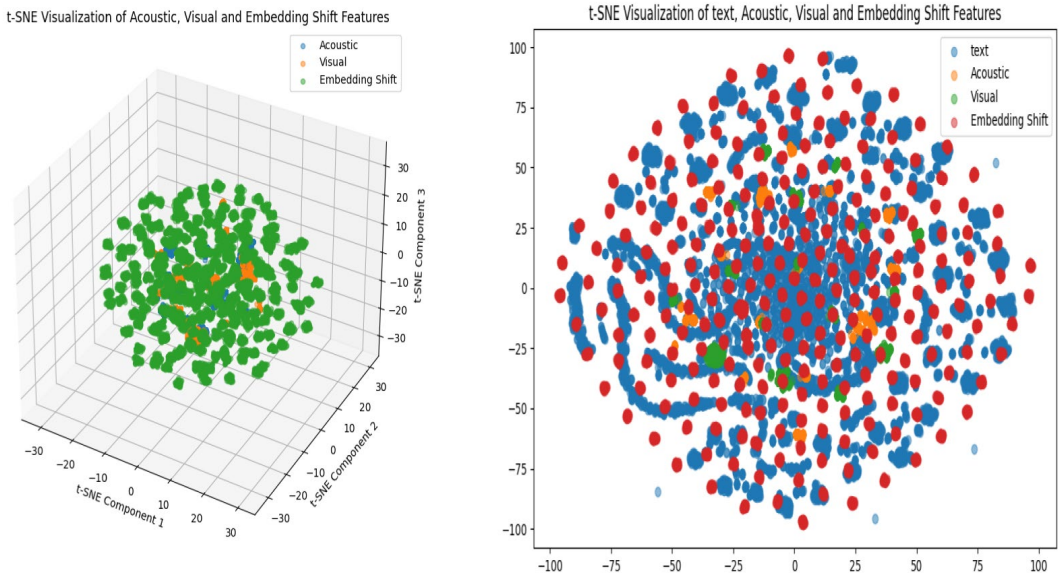
Model visualization

Features are visualized in dimensionality reduction. In Fig. 8, points of different colors represent different features, and their spatial distribution shows the similarities and differences between these features. For example, if points of the same color are clustered together, this indicates that these features are similar in high-dimensional space. First, Fig. 8a shows the dimensionality reduction of input text, video, and audio patterns after feature extraction. We can see that the text features occupy the widest area, which proves that the text information contains the richest emotional information. In Fig. 8b, the red dots (emotional feature embeddedness before PLM input) are clearly separated from the dots of other colors, which may indicate that the embeddedness offset feature is significantly different from the other features. Moreover, the distribution range of red dots includes points of other colors and, to a certain extent, points of other colors, indicating that the final emotional embedding can effectively integrate the emotional features of each mode.



(a) (b)

Fig. 7. Visualization of the average attention weight of the text guidance layer on the MOSI dataset.



(a) (b)

Fig. 8. Feature visualization of T-sne data after dimensionality reductiona.

Conclusions

In this paper, we propose a text guided multimodal scale path fusion network (TMSPF-Net). TMSPF-Net consists of three main modules: the Multi-scale Adaptive Transformer (MAT), the Text-Guided Conflict Elimination Module (TGCEM), and the sentiment channel fusion module. First, MAT fully extracts the more fine-grained emotional information through the combination of different scales, and collaborates the intra-modal attention and global emotional consistency through the dual attention mechanism. Meanwhile, the Mixture of Experts (MoE) is integrated in MAT to achieve dynamic path selection for efficient processing of heterogeneous data. Second, the Text-Guided Conflict Elimination Module (TGCEM) leverages hierarchical multi-scale text

embeddings to suppress contradictory signals in audio/visual modalities, enhancing their semantic alignment with linguistic cues. Finally our sentiment Optimization PLM extracts refined sentiment patterns from fused representations to improve the accuracy of multimodal sentiment analysis tasks. The experimental results on the dataset show that the TMSPF-Net has better performance than the current state-of-the-art (SOTA) model. The classification evaluation indicators on the MOSEI dataset are comprehensively higher than those of the SOTA model. The binary classification accuracy and seven-classification accuracy have increased by 0.37% and 0.88%, reaching 87.67% and 55.68% respectively. The seven-classification on the MOSI dataset has increased by 0.9% compared with the SOTA model and reached 50.5%. Meanwhile, the regression indicators of the TMSPF-Net model have been significantly improved compared with the SOTA model on both datasets. However, the following questions remain:

1. Overly complex models may be difficult to deploy in industry. 2. The development of large models is very rapid, providing new solutions for the alignment of text and video modalities.

Therefore, in our future design process, the computational complexity should be considered and a more lightweight model should be designed for deployment in actual engineering. We have observed that the MLP-based model shows advantages over the Transformer model in terms of lightweight and performance. Our next step is to consider applying the MLP-based model to the field of multimodal sentiment analysis and verify its effectiveness. Meanwhile, we plan to study the effectiveness of large models such as GPT in multimodal sentiment analysis tasks. The task of missing modalities is also one of our future research plans.

Data availability

The CMU-MOSI and CMU-MOSEI datasets, created and released by the Multicomp Lab at Carnegie Mellon University, support the findings of this study and are publicly available at: https://github.com/thuiar/Self-MM. Detailed data on the experimental process are available from Siyuan Liu (SYLiu@mail.sdu.edu.cn) upon reasonable request. The datasets are released under the MIT license, permitting use and publication. The images of the subjects in Figs. 1 and 4 are from CMU-MOSI, and all data in this dataset can be downloaded publicly. All subjects and/or their legal guardians have agreed to publish their identifying information or images in Scientific Reports after being fully informed.

Received: 25 June 2025; Accepted: 11 December 2025
Published online: 15 December 2025

References

1. Baltruvsaitis, T., Ahuja, C. & Morency, L.-P. Multimodal machine learning: A survey and taxonomy. *IEEE Trans. Pattern Anal. Mach. Intell.* **41**, 423–443 (2018).
2. Zadeh, A. et al. Memory fusion network for multi-view sequential learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32 (2018).
3. Zadeh, A., Chen, M., Poria, S., Cambria, E. & Morency, L.-P. Tensor fusion network for multimodal sentiment analysis. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing* 1103–1114 (Association for Computational Linguistics, 2017).
4. Liu, Z. et al. Efficient low-rank multimodal fusion with modality-specific factors. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* 2247–2256 (Association for Computational Linguistics, 2018).
5. Poria, S., Hazarika, D., Majumder, N. & Mihalcea, R. Beneath the tip of the iceberg: Current challenges and new directions in sentiment analysis research. *IEEE Trans. Affect. Comput.* **14**, 108–132 (2020).
6. Rahman, W. et al. Integrating multimodal information in large pretrained transformers. In*Proceedings of the Conference. Association for Computational Linguistics. Meeting*, vol. 2020, 2359 (2020).
7. Zhu, C. et al. Skeafn: Sentiment knowledge enhanced attention fusion network for multimodal sentiment analysis. *Inf. Fusion* **100**, 101958 (2023).
8. Wang, D. et al. Cross-modal enhancement network for multimodal sentiment analysis. *IEEE Trans. Multimedia* **25**, 4909–4921 (2023).
9. Hu, R., Yi, J., Chen, A. & Chen, L. Multichannel cross-modal fusion network for multimodal sentiment analysis considering language information enhancement. *IEEE Trans. Ind. Inf.* **20**, 9814–9824 (2024).
10. Zhong, M., Liu, Y., Xu, Y., Zhu, C. & Zeng, M. Dialoglm: Pre-trained model for long dialogue understanding and summarization. In *AAAI 2022* (2022).
11. Xia, Y. et al. A speaker-aware co-attention framework for medical dialogue information extraction. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing* (2022).
12. Yue, Y., Peng, Y. & Wang, D. Deep learning short text sentiment analysis based on improved particle swarm optimization. *Electronics* **12**, 23 (2023).
13. Wu, H. Multi-scale features enhanced sentiment region discovery for visual sentiment analysis. *Proc. SPIE* **12083**, 9 (2022).
14. Zadeh, A. et al. Multi-attention recurrent network for human communication comprehension. In *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32 (2018).
15. Tsai, Y.-H.H. et al. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the Conference. Association for Computational Linguistics. Meeting*, vol. 2019, 6558 (2019).
16. Shazeer, N. et al. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. Preprint at http://arXiv.org/1701.06538 (2017).
17. Han, Y. et al. Dynamic neural networks: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **44**, 7436–7456 (2021).
18. Wang, P., Yu, R., Gao, N., Lin, C. & Liu, Y. Task-driven data offloading for fog-enabled urban iot services. *IEEE Internet Things J.* **8**, 7562–7574 (2020).
19. Neverova, N., Wolf, C., Taylor, G. W. & Nebout, F. Multi-scale deep learning for gesture detection and localization. In *European Conference on Computer Vision* 474–490 (Springer, 2014).
20. Zhang, K., Gao, X., Tao, D. & Li, X. Single image super-resolution with multiscale similarity learning. *IEEE Trans. Neural Netw. Learn. Syst.* **24**, 1648–1659 (2013).
21. Lin, T.-Y. et al. Feature pyramid networks for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 2117–2125 (2017).
22. Gu, Y. et al. Multimodal affective analysis using hierarchical attention strategy with word-level alignment. In *Proceedings of the Conference. Association for Computational Linguistics. Meeting*, vol. 2018, 2225 (2018).

23. Lei, Y., Yang, S., Wang, X. & Xie, L. Msemotts: Multi-scale emotion transfer, prediction, and control for emotional speech synthesis. *IEEE/ACM Trans. Audio Speech Lang. Process.* **30**, 853–864 (2022).
24. Cao, X., Liangwen, H., Wang, H. & Liu, L. Microblog-oriented multi-scale cnn multi-label sentiment classification model. In *2020 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT)* 626–631 (2020).
25. Lin, C., Cheng, H., Rao, Q. & Yang, Y. M3sa: Multimodal sentiment analysis based on multi-scale feature extraction and multi-task learning. *IEEE/ACM Trans. Audio Speech Lang. Process.* **32**, 1416–1429 (2024).
26. Fu, J., Fu, Y., Xue, H. & Xu, Z. Tmfn: a text-based multimodal fusion network with multi-scale feature extraction and unsupervised contrastive learning for multimodal sentiment analysis. *Complex Intell. Syst.* **11**, 1–16 (2025).
27. Sabour, S., Frosst, N. & Hinton, G. E. Dynamic routing between capsules. *Adv. Neural Inf. Process. Syst.* **30**, 1 (2017).
28. Hazarika, D., Zimmermann, R. & Poria, S. Misa: Modality-invariant and-specific representations for multimodal sentiment analysis. In *Proceedings of the 28th ACM International Conference on Multimedia* 1122–1131 (2020).
29. He, J., Mai, S. & Hu, H. A unimodal reinforced transformer with time squeeze fusion for multimodal sentiment analysis. *IEEE Signal Process. Lett.* **28**, 992–996 (2021).
30. Sun, X., He, H., Tang, H., Zeng, K. & Shen, T. Multimodal rough set transformer for sentiment analysis and emotion recognition. In *2023 IEEE 9th International Conference on Cloud Computing and Intelligent Systems (CCIS)* 250–259 (IEEE, 2023).
31. Huan, R., Zhong, G., Chen, P. & Liang, R. Unimf: A unified multimodal framework for multimodal sentiment analysis in missing modalities and unaligned multimodal sequences. *IEEE Trans. Multimedia* **26**, 5753–5768 (2023).
32. Huan, R., Zhong, G., Chen, P. & Liang, R. Trisat: Trimodal representation learning for multimodal sentiment analysis. *IEEE/ACM Trans. Audio Speech Lang. Process.* **32**, 4105–4120 (2024).
33. Lu, J., Batra, D., Parikh, D. & Lee, S. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Adv. Neural Inf. Process. Syst.* **32**, 1 (2019).
34. Su, W. et al. Vl-bert: Pre-training of generic visual-linguistic representations. In *International Conference on Learning Representations* (2020).
35. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, volume 1 (Long and Short Papers)* 4171–4186 (2019).
36. Zhuang, L., Wayne, L., Ya, S. & Jun, Z. A robustly optimized bert pre-training approach with post-training. In *Proceedings of the 20th Chinese National Conference on Computational Linguistics* 1218–1227 (Chinese Information Processing Society of China, 2021).
37. Ke, P., Ji, H., Liu, S., Zhu, X. & Huang, M. Sentilare: Sentiment-aware language representation learning with linguistic knowledge. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* 6975–6988 (Association for Computational Linguistics, 2020).
38. Degottex, G., Kane, J., Drugman, T., Raitio, T. & Scherer, S. *Covarep: A Collaborative Voice Analysis Repository for Speech Technologies* (IEEE, 2014).
39. Amos, B., Ludwiczuk, B. & Satyanarayanan, M. Openface: A general-purpose face recognition library with mobile applications. *Tech. Rep., CMU-CS-16-118, CMU School of Computer Science* (2016).
40. Zadeh, A., Zellers, R., Pincus, E. & Morency, L.-P. Mosi: Multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos. Preprint at http://arXiv.org/1606.06259 (2016).
41. Bagher Zadeh, A., Liang, P. P., Poria, S., Cambria, E. & Morency, L.-P. Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* 2236–2246 (2018).
42. Yang, Z., He, Q., Du, N. & He, Q. Temporal text-guided feedback-based progressive fusion network for multimodal sentiment analysis. *Alex. Eng. J.* **116**, 699–709 (2025).
43. Zhang, H. et al. Learning language-guided adaptive hyper-modality representation for multimodal sentiment analysis. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (eds Bouamor, H. et al.), 756–767 (Association for Computational Linguistics, 2023).
44. Guo, X., Tian, S., Yu, L. & He, X. Smartran: Smart routing attention network for multimodal sentiment analysis. *Appl. Intell.* **54**, 12742–12763 (2024).
45. Zhao, X., Wang, M., Tan, Y. & Wang, X. Tgmoe: A text guided mixture-of-experts model for multimodal sentiment analysis. *Int. J. Adv. Comput. Sci. Appl.* **15**, 1 (2024).
46. Hu, C., Liu, J., Li, X., Li, M. & He, H. Mst-argcn: modality-squeeze transformer with attentional recurrent graph capsule network for multimodal sentiment analysis. *J. Supercomput.* **81**, 86 (2025).
47. Liu, C., Wang, Y. & Yang, J. A transformer-encoder-based multimodal multi-attention fusion network for sentiment analysis. *Appl. Intell.* **54**, 8415–8441 (2024).
48. Wang, L., Peng, J., Zheng, C., Zhao, T. & Zhu, L. A cross modal hierarchical fusion multimodal sentiment analysis method based on multi-task learning. *Inf. Process. Manag.* **61**, 103675 (2024).
49. Huang, J., Zhou, J., Tang, Z., Lin, J. & Chen, C.Y.-C. Tmbl: Transformer-based multimodal binding learning model for multimodal sentiment analysis. *Knowl.-Based Syst.* **285**, 111346 (2024).
50. Lin, H. et al. Ps-mixer: A polar-vector and strength-vector mixer model for multimodal sentiment analysis. *Inf. Process. Manag.* **60**, 103229 (2023).
51. Huang, J. et al. Dynamic hypergraph convolutional network for multimodal sentiment analysis. *Neurocomputing* **565**, 126992 (2024).
52. Wang, B. et al. Tripartite interaction representation learning for multi-modal sentiment analysis. *Expert Syst. Appl.* **268**, 126279 (2025).
53. Wang, B. et al. Non-uniform attention network for multi-modal sentiment analysis. In *MultiMedia Modeling* 612–623 (2022).

Author contributions

S.L.: Conceptualization, Methodology, Validation, Investigation, Writing-Original Draft, Writing-Review & Editing, Visualization. H.Z.: Validation, Writing-Review & Editing, Supervision. Y.C.: Validation, Writing-Review & Editing, Supervision. F.K: Validation, Writing-Review & Editing, Supervision. K.L.: Validation, Writing-Review & Editing, Supervision, Funding Acquisition.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to K.L.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.