Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Register-Augmented Attention and Self-Calibrated Fusion for Robust Multimodal Sentiment Analysis

Bin Xu
Zhejiang Gongshang University
Weiyang Wang
Zhejiang Gongshang University
Ao Ding
dingao@pop.zjgsu.edu.cn

Zhejiang Gongshang University

Research Article

Keywords: Multimodal sentiment analysis, Register-augmented attention, Self-calibrated fusion, Fine-grained sentiment classification

Posted Date: April 29th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9134197/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Version of Record: A version of this preprint was published at Multimedia Systems on September 6th, 2026. See the published version at https://doi.org/10.1007/s00530-026-02653-8.

Register-Augmented Attention and Self-Calibrated Fusion for Robust Multimodal Sentiment Analysis

Bin Xu[1], Weiyang Wang[1], Ao Ding[1*]

[1*]School of Information and Electronic Engineering, Zhejiang Gongshang University, 18 Xuezheng Street, Hangzhou, 310018, China.

*Corresponding author(s). E-mail(s): dingao@pop.zjgsu.edu.cn;
Contributing authors: xubin@zjgsu.edu.cn; weiyang.wang@pop.zjgsu.edu.cn;

Abstract

Multimodal sentiment analysis (MSA) aims to understand human emotions by integrating heterogeneous modalities such as language and acoustic signals. Despite recent progress, two key challenges remain: (1) inter-modal inconsistency, where different modalities may convey conflicting sentiment cues, and (2) intra-modal feature ambiguity caused by noise and subtle emotional variations.
To address these issues, we propose RegCal-Net, a register-augmented and self-calibrated framework for bimodal sentiment analysis. First, we introduce a Register-Augmented Self-Attention (RASA) mechanism that incorporates learnable register tokens to provide stable global anchors and absorb redundant information within each modality. Second, we design a Self-Calibrated Fusion (SCF) module that dynamically evaluates fused feature reliability through a prototype-guided gating strategy, enabling adaptive suppression of unreliable signals during multimodal integration.
Extensive experiments on two widely used benchmarks, CMU-MOSI and CMU-MOSEI, demonstrate that RegCal-Net achieves competitive performance across multiple evaluation metrics, particularly improving fine-grained sentiment classification and reducing prediction error. These results indicate that combining register-based representation stabilization with quality-aware fusion provides an effective solution for robust multimodal sentiment analysis.

Keywords: Multimodal sentiment analysis, Register-augmented attention, Self-calibrated fusion, Fine-grained sentiment classification

1 Introduction

Sentiment analysis aims to identify and comprehend human attitudes, opinions, and emotions from data. While traditional methods primarily focus on textual information, human communication is inherently multimodal, involving acoustic tones, facial expressions, and body language [1]. Consequently, multimodal sentiment analysis (MSA) has emerged as a vibrant research field, with applications in social media monitoring, human-computer interaction, and healthcare diagnosis [2, 3].

Despite considerable progress, effectively fusing multimodal data remains challenging due to the "heterogeneity gap." Different modalities exhibit distinct statistical properties and semantic densities. For instance, text typically conveys dense semantic meaning, while audio provides contextual intensity (with video integration reserved for future work). Recent works such as MulT [4]

and MISA [5] have attempted to bridge this gap, but two critical issues persist:

(1) **Inter-modal inconsistency and noise**: Modalities do not always align. In sarcasm or irony, text may express positive sentiment while audio tone is negative. Naive fusion strategies often average these conflicting signals, leading to erroneous predictions. Moreover, noise in individual modalities can degrade the fused representation quality.

(2) **Intra-modal feature ambiguity**: In fine-grained sentiment classification, boundaries between adjacent scores are often blurred. Standard self-attention mechanisms tend to mix features globally, potentially smoothing out subtle discriminative features.

To address these limitations, we propose RegCal-Net. We introduce the Register-Augmented Self-Attention (RASA) mechanism, inspired by Vision Transformers, which uses register tokens as sinks for redundant information. Furthermore, we propose the Self-Calibrated Fusion (SCF) module, which employs a prototype-based scoring system to dynamically admit or suppress information, thereby filtering out conflicting signals.

Our main contributions are summarized as follows:

1. We identify two critical yet under-explored challenges in bimodal sentiment analysis: representation instability caused by intra-modal noise and unreliable multimodal fusion under conflicting modality cues.

2. To address representation instability, we propose a Register-Augmented Self-Attention (RASA) mechanism that introduces learnable register tokens as global anchors to stabilize attention patterns and absorb redundant information within each modality.

3. To improve multimodal integration reliability, we design a Self-Calibrated Fusion (SCF) module that dynamically evaluates fused feature quality using a prototype-guided gating strategy.

4. Extensive experiments on the CMU-MOSI and CMU-MOSEI benchmarks demonstrate that the proposed framework consistently improves fine-grained sentiment prediction performance while reducing regression errors.

2 Related Work

2.1 Multimodal Sentiment Analysis

Multimodal sentiment analysis has evolved from simple early-fusion approaches to sophisticated architectures that explicitly model cross-modal interactions. Tensor Fusion Network (TFN) [6] introduced tensor products for capturing multimodal interactions, while Low-rank Multimodal Fusion (LMF) [7] addressed computational efficiency through low-rank approximations. Attention-based methods, such as Multimodal Transformer (MulT) [4], leverage cross-modal attention to learn aligned representations. More recent approaches focus on modality-specific and modality-invariant feature decomposition [5] or introduce contrastive learning strategies [8] to strengthen multimodal fusion. However, these methods often overlook the stability of learned representations and lack explicit mechanisms for assessing fusion quality, which our RegCal-Net addresses through register augmentation and self-calibration. Recent works such as AtCAF [9], UCMIB-PNS [10], H2CAN [11], PAMoE-MSA [12], and Actual Cause Guided AGS [13] further advance causal and uncertainty-aware fusion but still lack stable intra-modal anchors and online quality self-assessment.

Recent studies have further explored reliability-aware and causality-inspired multimodal fusion strategies. For instance, causality-aware fusion methods attempt to model modality interactions from a causal perspective to improve robustness under modality conflicts. Other works introduce uncertainty estimation or modality reliability modeling to dynamically adjust modality contributions during fusion. While these approaches improve robustness, many of them rely on complex causal modeling or additional supervision signals. In contrast, our approach focuses on improving representation stability through register-augmented attention and introducing a lightweight self-calibration mechanism that dynamically evaluates fusion quality during inference.

The MPC model constructs a highly discriminative textual feature space using multi-layer progressive contrastive learning with hierarchically structured hard positive samples from different BERT layers and progressively challenging

hard negative samples from other modalities. It then applies a conditional transform-based feature enhancement module to recalibrate non-textual modalities and a sentiment-aware fusion module for weighted integration. While MPC achieves strong improvements on small datasets by mitigating modality imbalance, it relies on static textual anchors and does not provide explicit mechanisms for intra-modal noise absorption via register tokens or dynamic online quality assessment during fusion. RegCal-Net complements and advances this line of work by introducing stable global register anchors that absorb redundant information within each modality and prototype-guided dynamic gating that filters unreliable fused signals in real time, leading to superior fine-grained performance and lower MAE.

2.2 Attention with Registers

The self-attention mechanism, introduced by Vaswani et al. [14], is fundamental to modern sequence modeling. To enhance model capacity, recent works have explored adding learnable auxiliary tokens. A prominent example is register tokens, where Darcet et al. [15] proposed their inclusion in Vision Transformers to serve as global information aggregators. These registers stabilize attention patterns by absorbing redundant information. Our work extends this concept to MSA for the first time, using register tokens in our RASA mechanism to mitigate feature ambiguity within each modality.

2.3 Calibration and Gating

To improve model reliability and control information flow during fusion, calibration and gating mechanisms are widely used. Gating, which allows adaptive control of information, has a long history in neural networks. For example, Hochreiter and Schmidhuber [16] introduced gating units as a core component of the Long Short-Term Memory architecture to manage the cell state.

In the context of multimodal fusion, gating has been applied to selectively integrate features. For example, Wang et al. [17] proposed an adaptive gating mechanism to dynamically adjust the contribution of different feature sources in multimodal tasks. However, existing calibration methods in MSA are typically applied post hoc, and gating mechanisms often lack an explicit signal to judge the quality or reliability of the features being fused. Our SCF module introduces a prototype-score-guided gating approach that performs online calibration during the fusion process itself. It dynamically assesses feature quality based on a learned sentiment prototype, enabling more robust and reliable integration of multimodal information.

3 Proposed Method

3.1 Problem Formulation

We focus on bimodal sentiment analysis using text and audio. For a target utterance, let $\mathbf{X}_t \in \mathbb{R}^{l_t \times d_t}$ and $\mathbf{X}_a \in \mathbb{R}^{l_a \times d_a}$ denote its text and audio inputs, where l_m and d_m are the sequence length and feature dimension for modality $m \in \{t, a\}$.

In addition to the target utterance, we incorporate a context sample from the same video segment. We denote the corresponding pre-constructed context-level inputs by $\mathbf{X}_{\text{ctx},t}$ and $\mathbf{X}_{\text{ctx},a}$ for text and audio, respectively. These context inputs are encoded using the same text and audio encoders as the target utterance.

Given the target-level and context-level bimodal inputs, the goal of the model is to predict both a continuous sentiment score $\tilde{y} \in \mathbb{R}$ and a discrete sentiment category $\tilde{c} \in \{1, \ldots, C\}$.

3.2 Overall Framework

The overall architecture of RegCal-Net is shown in Fig. 1. The framework contains three stages:

1. **Feature extraction:** a pre-trained RoBERTa-large encoder is used for text and a pre-trained Data2Vec-Audio-large encoder is used for audio.
2. **Register-augmented cross-modal encoding:** multiple RA-CME layers perform cross-modal interaction, register-augmented self-attention, and feed-forward transformation.
3. **Self-calibrated fusion:** the final [CLS] representations from target and context streams are fused, modulated by an auxiliary scalar score, and used for sentiment regression and classification.

The model processes four streams in parallel: target-text, target-audio, context-text, and context-audio. For both text and audio streams,

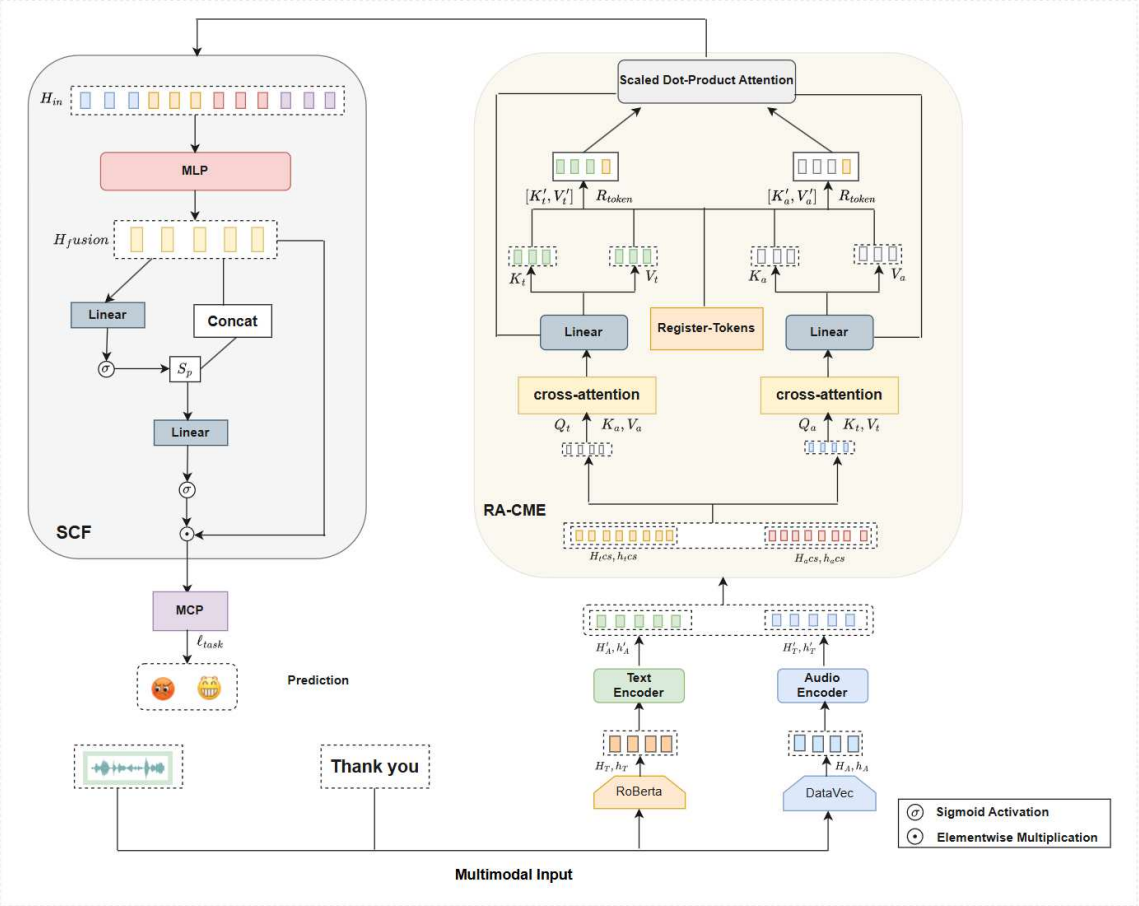


Fig. 1 The overall architecture of RegCal-Net (bimodal text-audio setting). Context-level features are obtained via parallel global average pooling.

a learnable [CLS] embedding is prepended before entering the RA-CME layers. At each RA-CME layer, a scalar gate is predicted from the current text and audio [CLS] features, and this gate interpolates between the newly updated features and the previous features. After the stacked RA-CME layers, the [CLS] features from all four streams are fused for downstream prediction.

3.3 Feature Extraction

For text, we use RoBERTa-large to obtain the hidden sequence representation:

$$\mathbf{H}_t = \mathrm{Enc}_t(\mathbf{X}_t), \qquad \mathbf{H}_{\mathrm{ctx},t} = \mathrm{Enc}_t(\mathbf{X}_{\mathrm{ctx},t}). \quad (1)$$

For audio, we use Data2Vec-Audio-large to extract hidden acoustic features:

$$\mathbf{H}_a = \mathrm{Enc}_a(\mathbf{X}_a), \qquad \mathbf{H}_{\mathrm{ctx},a} = \mathrm{Enc}_a(\mathbf{X}_{\mathrm{ctx},a}). \quad (2)$$

In addition to sequence-level hidden states, pooled representations from the text and audio encoders are also used for auxiliary unimodal predictions in training.

Before the multimodal encoder, a learnable [CLS] embedding is prepended to each target/context text/audio sequence:

$$\widehat{\mathbf{H}} = [\mathbf{e}_{\mathrm{cls}}; \mathbf{H}], \quad (3)$$

where $\mathbf{e}_{\mathrm{cls}} \in \mathbb{R}^d$ is a trainable modality-specific [CLS] embedding.

3.4 Register-Augmented Attention

3.4.1 RA Self-Attention

Given an input sequence $\mathbf{X} \in \mathbb{R}^{B\times L\times d}$, we compute the query, key, and value tensors:

$$\mathbf{Q} = \mathbf{X}\mathbf{W}_Q, \quad \mathbf{K} = \mathbf{X}\mathbf{W}_K, \quad \mathbf{V} = \mathbf{X}\mathbf{W}_V, \quad (4)$$

where $\mathbf{W}_Q, \mathbf{W}_K, \mathbf{W}_V \in \mathbb{R}^{d\times d}$. We then split the hidden dimension into H heads with per-head dimension $d_h = d/H$, yielding

$$\mathbf{Q}, \mathbf{K}, \mathbf{V} \in \mathbb{R}^{B\times H\times L\times d_h}. \quad (5)$$

The standard scaled dot-product attention logits are computed as

$$\mathbf{A} = \frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_h}}, \qquad \mathbf{A} \in \mathbb{R}^{B\times H\times L\times L}. \quad (6)$$

To augment self-attention, we introduce R learnable register heads. In our implementation, the register attention logits and register values are parameterized by two learnable tensors:

$$\mathbf{R}^A \in \mathbb{R}^{R\times L_{\max}\times L_{\max}}, \qquad \mathbf{R}^V \in \mathbb{R}^{R\times d_h\times L_{\max}}, \quad (7)$$

where $L_{\max}$ is the maximum supported sequence length. For an input sequence of current length L, these tensors are sliced to

$$\mathbf{R}^A_L \in \mathbb{R}^{R\times L\times L}, \qquad \mathbf{R}^V_L \in \mathbb{R}^{R\times d_h\times L}. \quad (8)$$

After reshaping and broadcasting along the batch dimension, the register tensors become

$$\mathbf{R}^A_L \in \mathbb{R}^{B\times R\times L\times L}, \qquad \mathbf{R}^V_L \in \mathbb{R}^{B\times R\times L\times d_h}. \quad (9)$$

We concatenate the register tensors with the standard attention logits and values along the head dimension:

$$\widetilde{\mathbf{A}} = \mathrm{Concat}_{\mathrm{head}}(\mathbf{A}, \mathbf{R}^A_L) \in \mathbb{R}^{B\times(H+R)\times L\times L}, \quad (10)$$

$$\widetilde{\mathbf{V}} = \mathrm{Concat}_{\mathrm{head}}(\mathbf{V}, \mathbf{R}^V_L) \in \mathbb{R}^{B\times(H+R)\times L\times d_h}. \quad (11)$$

The register-augmented attention output is then computed as

$$\widetilde{\mathbf{C}} = \mathrm{Softmax}(\widetilde{\mathbf{A}})\widetilde{\mathbf{V}}, \qquad \widetilde{\mathbf{C}} \in \mathbb{R}^{B\times(H+R)\times L\times d_h}. \quad (12)$$

Following the implementation, only the first H channels corresponding to the original attention heads are retained:

$$\mathbf{C} = \widetilde{\mathbf{C}}^{(1:H)} \in \mathbb{R}^{B\times H\times L\times d_h}. \quad (13)$$

After merging the head dimension back to $\mathbb{R}^{B\times L\times d}$, we apply a residual connection and layer normalization:

$$\mathbf{H} = \mathrm{LayerNorm}(\mathrm{MergeHeads}(\mathbf{C}) + \mathbf{X}). \quad (14)$$

3.4.2 Cross-Modal Attention

To model direct interaction between modalities, we use cross-modal attention. For example, text-to-audio attention is defined as

$$\mathbf{O}^{(t)} = \mathrm{Softmax}\left(\frac{\mathbf{Q}^{(t)}(\mathbf{K}^{(a)})^\top}{\sqrt{d_h}}\right)\mathbf{V}^{(a)}, \quad (15)$$

and audio-to-text attention is computed symmetrically.

3.4.3 RA-CME Layer

Each RA-CME layer consists of three steps:

1. **Cross-modal attention:** text attends to audio and audio attends to text.
2. **Register-augmented self-attention:** each modality is further refined by the proposed RASA module.
3. **Feed-forward transformation:** modality-specific feed-forward networks update the refined features.

In addition, each RA-CME layer uses a scalar gate to control the update intensity. Let $\mathbf{h}^{(i)}_{t,\mathrm{cls}}$ and $\mathbf{h}^{(i)}_{a,\mathrm{cls}}$ denote the text and audio [CLS] features at the input of the i-th layer. A scalar gate is predicted as

$$s^{(i)} = \sigma\left(f_g\left([\mathbf{h}^{(i)}_{t,\mathrm{cls}}; \mathbf{h}^{(i)}_{a,\mathrm{cls}}]\right)\right), \quad (16)$$

where $f_g(\cdot)$ is a lightweight MLP and $s^{(i)} \in (0,1)$. Let $\hat{\mathbf{X}}_t^{(i)}$ and $\hat{\mathbf{X}}_a^{(i)}$ denote the outputs of the current RA-CME block. The final update is computed by

$$\mathbf{X}_t^{(i+1)} = s^{(i)}\hat{\mathbf{X}}_t^{(i)} + (1 - s^{(i)})\mathbf{X}_t^{(i)}, \tag{17}$$

$$\mathbf{X}_a^{(i+1)} = s^{(i)}\hat{\mathbf{X}}_a^{(i)} + (1 - s^{(i)})\mathbf{X}_a^{(i)}. \tag{18}$$

The same procedure is applied to the context streams.

3.5 Self-Calibrated Fusion Module

After the stacked RA-CME layers, we obtain the final output sequences for the target and context streams. We extract the representation at the first position, corresponding to the prepended [CLS] token:

$$\mathbf{h}_t = \mathbf{X}_t^{(N)}[:,0], \qquad \mathbf{h}_a = \mathbf{X}_a^{(N)}[:,0], \tag{19}$$

$$\mathbf{h}_{\text{ctx},t} = \mathbf{X}_{\text{ctx},t}^{(N)}[:,0], \quad \mathbf{h}_{\text{ctx},a} = \mathbf{X}_{\text{ctx},a}^{(N)}[:,0]. \tag{20}$$

We concatenate the four [CLS] features and pass them through a two-layer fusion network:

$$\mathbf{h}_{\text{in}} = [\mathbf{h}_t; \mathbf{h}_a; \mathbf{h}_{\text{ctx},t}; \mathbf{h}_{\text{ctx},a}], \tag{21}$$

$$\mathbf{h}_{\text{fused}} = \text{MLP}(\mathbf{h}_{\text{in}}). \tag{22}$$

From the fused representation, we further predict an auxiliary scalar score:

$$s_p = \mathbf{w}_p^\top \mathbf{h}_{\text{fused}} + b_p. \tag{23}$$

We then concatenate this scalar score with the fused feature and produce a feature-wise gate:

$$\mathbf{g}_{\text{in}} = [\mathbf{h}_{\text{fused}}; s_p], \tag{24}$$

$$\mathbf{g} = \sigma(\mathbf{W}_g \mathbf{g}_{\text{in}} + \mathbf{b}_g), \qquad \mathbf{h}_{\text{gate}} = \mathbf{h}_{\text{fused}} \odot \mathbf{g}. \tag{25}$$

Finally, $\mathbf{h}_{\text{gate}}$ is fed into two output heads: a regression head for continuous sentiment prediction and a classification head for discrete sentiment recognition.

3.6 Auxiliary Unimodal Prediction

Besides the main multimodal prediction, we also introduce auxiliary unimodal predictions from text and audio. Specifically, the pooled target/context text features are concatenated to form a text-side auxiliary representation, and the pooled target/context audio features are concatenated to form an audio-side auxiliary representation:

$$\mathbf{z}_t = [\mathbf{p}_t; \mathbf{p}_{\text{ctx},t}], \tag{26}$$

$$\mathbf{z}_a = [\mathbf{p}_a; \mathbf{p}_{\text{ctx},a}], \tag{27}$$

where $\mathbf{p}_t, \mathbf{p}_{\text{ctx},t}, \mathbf{p}_a, \mathbf{p}_{\text{ctx},a}$ are encoder-level pooled outputs. These auxiliary branches are used to produce text-only and audio-only predictions during training.

3.7 Training Objective

The model is trained with a multi-task objective consisting of classification loss, regression loss, and auxiliary score supervision. Let $\hat{y}$ denote the final regression output, $\hat{\mathbf{c}}$ the classification logits, and s_p the auxiliary scalar score predicted from the fused representation. The regression loss is

$$\mathcal{L}_{\text{reg}} = \text{MSE}(\hat{y}, y) + \alpha\, \mathcal{L}_{\text{corr}}, \tag{28}$$

where $\mathcal{L}_{\text{corr}}$ is the Pearson-correlation-based loss:

$$\mathcal{L}_{\text{corr}} = 1 - \frac{\sum_i (y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{\sqrt{\sum_i (y_i - \bar{y})^2}\sqrt{\sum_i (\hat{y}_i - \bar{\hat{y}})^2}}. \tag{29}$$

For sentiment classification, we use cross-entropy loss:

$$\mathcal{L}_{\text{cls}} = \text{CE}(\hat{\mathbf{c}}, c). \tag{30}$$

To encourage the auxiliary scalar score to encode sentiment-related global information, we supervise it using the sentiment label:

$$\mathcal{L}_{\text{aux}} = \text{MSE}(s_p, y). \tag{31}$$

The total loss is

$$\mathcal{L}_{\text{total}} = \lambda_{\text{cls}}\mathcal{L}_{\text{cls}} + \lambda_{\text{reg}}\mathcal{L}_{\text{reg}} + \lambda_{\text{aux}}\mathcal{L}_{\text{aux}}. \tag{32}$$

In our implementation, the default loss weights are set to

$$\lambda_{\text{cls}} = 1.0, \qquad \lambda_{\text{reg}} = 0.2, \qquad \lambda_{\text{aux}} = 0.1. \tag{33}$$

Table 1 Dataset statistics.

Dataset	Train	Valid	Test	Total
CMU-MOSI	1284	296	686	2199
CMU-MOSEI	16326	1871	4659	22856

The classification branch uses label smoothing, while the regression branch additionally includes the correlation-based regularization term.

4 Experiments

4.1 Experiment Settings

Although CMU-MOSI and CMU-MOSEI provide visual modalities, this work focuses on the text-audio bimodal setting because linguistic and acoustic signals constitute the primary channels for expressing sentiment in many real-world conversational scenarios. Moreover, isolating the bimodal setting allows clearer analysis of cross-modal interaction mechanisms without introducing additional visual noise. Extending the framework to tri-modal scenarios will be explored in future work. To rigorously assess the efficacy of our proposed approach, we conducted comprehensive experiments and ablation studies on the standard CMU-MOSI and CMU-MOSEI benchmarks. These datasets consist of annotated movie review video segments with sentiment scores reflecting both polarity and intensity. Evaluation metrics include binary accuracy (Acc-2), 5-class accuracy (Acc-5), 7-class accuracy (Acc-7), F1 score, Mean Absolute Error (MAE), and Pearson correlation (Corr). For binary metrics, we adopt a negative/non-negative split for MOSI and a negative/positive split for MOSEI.

4.1.1 Datasets and Evaluation Metrics

CMU-MOSI [18] contains 2,199 opinion video clips from 93 YouTube movie review videos, each annotated with sentiment intensity scores in $[-3, +3]$ and discrete labels for seven-level classification.

CMU-MOSEI [19] is a larger-scale dataset with 23,453 annotated video segments from 1,000 speakers and 250 topics, providing both regression scores and classification labels.

Evaluation Metrics: Following standard protocols, we report:

- Binary accuracy (Acc-2): accuracy for negative vs. positive classification.
- Seven-class accuracy (Acc-7): fine-grained sentiment classification accuracy.
- F1 score: harmonic mean of precision and recall for binary classification.
- Mean Absolute Error (MAE): average absolute difference between predicted and true regression scores.
- Pearson correlation (Corr): linear correlation between predictions and ground truth.

4.1.2 Implementation Details

Feature Extraction: We use pre-trained RoBERTa-large [20] for text encoding and Data2Vec-Audio-large for audio encoding, both from HuggingFace Transformers.

Model Configuration: Our RegCal-Net uses $N = 3$ RA-CME layers with $H = 8$ attention heads and $d = 768$ hidden dimension. Register tokens use $R = 4$ heads. The SCF fusion network uses intermediate dimension $2d = 1536$.

Training: We train with AdamW optimizer (learning rate 1×10^{-5}, weight decay 0.01), batch size 16, for 24 epochs. The loss weights are set to $\lambda_{\text{cls}} = 1.0$, $\lambda_{\text{reg}} = 0.2$, and $\lambda_{\text{aux}} = 0.1$. Dropout rate is 0.1.

Hardware: All experiments are conducted on NVIDIA RTX4090 GPUs with 24GB memory.

4.2 Performance Comparison

We compare RegCal-Net with several strong baselines and recent state-of-the-art methods on both datasets. As shown in Table 2, RegCal-Net achieves superior performance across most metrics on both datasets.

4.3 Ablation Study and Analysis

To validate the contribution of each component, we conduct comprehensive ablation studies on the CMU-MOSI dataset. The results are presented in Table 3.

4.3.1 Impact of Register Count

To further investigate the impact of register tokens, we conduct sensitivity experiments by varying the number of register heads R. As shown in Table 4, performance improves markedly when

Table 2 Performance comparison on CMU-MOSI and CMU-MOSEI datasets.

Method	MOSI						MOSEI					
	Acc-7	Acc-5	Acc-2	F1	MAE↓	Corr↑	Acc-7	Acc-5	Acc-2	F1	MAE↓	Corr↑
TFN [6]	36.15	40.52	77.99/78.81	77.92/78.78	0.9319	0.6750	51.32	52.80	80.85/83.65	81.13/83.43	0.5770	0.7147
MISA [5]	43.44	49.56	81.78/83.38	81.78/83.44	0.7531	0.7819	52.05	53.63	80.67/84.67	81.12/84.66	0.5576	0.7515
Self-MM [8]	45.92	52.92	83.09/84.60	83.04/84.06	0.7159	0.7919	53.87	55.53	83.76/85.15	83.82/84.90	0.5309	0.7649
ALMT [21]	46.15	52.31	83.22/85.38	82.29/84.39	0.7012	0.7909	52.52	54.26	82.36/85.58	81.95/85.63	0.5323	0.7751
TETFN [22]	45.77	53.64	83.24/85.37	83.13/85.33	0.7084	0.7984	53.90	55.78	84.12/86.21	84.35/86.11	0.5373	0.7696
LMF [7]	36.01	42.13	77.70/78.66	77.72/78.75	0.9458	0.6600	53.29	54.78	79.95/83.90	80.52/83.88	0.5605	0.7356
MFN [23]	34.84	39.65	78.13/79.42	78.11/79.46	0.9471	0.6534	52.31	53.81	81.05/84.12	81.44/84.01	0.5655	0.7257
GRAPH-MFN [24]	33.24	37.32	76.53/77.44	76.42/77.38	0.9719	0.6582	52.24	53.60	77.40/83.38	78.28/83.48	0.5569	0.7321
MulT [4]	34.69	40.52	79.15/81.10	78.83/80.87	0.9125	0.6783	52.84	54.18	81.15/84.63	81.56/84.52	0.5593	0.7331
MAG-BERT [25]	44.75	50.29	81.92/83.23	81.89/83.25	0.7408	0.7831	52.78	54.54	82.21/85.25	85.62/85.21	0.5453	0.7556
FDMER[a] [26]	44.10	–	–/84.60	–/83.90	0.7240	0.7730	54.10	–	–/86.10	–/85.80	0.5360	0.7730
MMIM[a] [27]	46.65	–	84.14/86.06	84.00/85.98	0.7000	0.8000	54.24	–	82.24/85.97	82.66/85.94	0.5260	0.7720
Hycon[a] [28]	46.60	–	–/85.20	–/85.10	0.7130	0.7900	52.80	–	–/85.40	–/85.60	0.6010	0.7760
MMML[b] [29]	50.34	58.16	87.51/89.69	87.45/89.67	0.5830	0.8690	55.74	58.57	**87.24**/88.02	**87.18**/88.15	0.4920	0.8140
RegCal-Net	**51.90**	**60.50**	**88.19/90.55**	**88.10/90.51**	**0.5545**	**0.8848**	**56.99**	**58.83**	85.79/**88.06**	85.91/87.90	**0.4875**	**0.8166**

Note: The best result is highlighted in bold.
In Acc-2 and F1 scores, the left side of "/" is calculated as "negative/non-negative", and the right side is calculated as "positive/negative".
[a] These results are from [26–28].
[b] Represents that the result is from [29].

Table 3 Ablation study showing the effects of different components.

No.	Model Variant	Acc-5	MAE↓
0	**RegCal-Net (Full)**	**60.50**	**0.5545**
1	w/o Register Tokens	59.62	0.5768
2	w/o SCF	59.62	0.5537

increasing R from 1 to 4, achieving the highest Acc-2 (0.8746), F1 (0.8743), lowest MAE (0.5735), and highest Corr (0.8640). This confirms that four register heads provide an optimal balance: too few registers ($R = 1$) fail to sufficiently absorb redundant intra-modal noise, while too many ($R = 8$) introduce unnecessary parameters that slightly degrade generalization. These results directly validate the effectiveness of our RASA mechanism in stabilizing representations for fine-grained sentiment discrimination.

Table 4 Impact of register count R on MOSI (5-run mean±std).

R	Acc-2	F1	MAE↓	Corr↑
1	0.8586	0.8581	0.5910	0.8419
4	**0.8746**	**0.8743**	**0.5735**	**0.8640**
8	0.8688	0.8681	0.5877	0.8565

4.3.2 Impact of RA-CME Layers

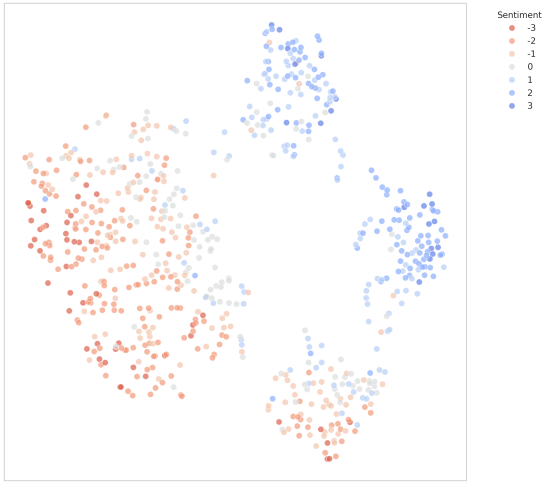
We also examine the sensitivity with respect to the number of RA-CME layers N. As reported in Table 5, model performance steadily improves as N increases from 1 to 3, reaching the best Acc-2 (87.17), F1 (87.16), MAE (0.5810), and Corr (0.8610) at $N = 3$. This trend demonstrates that deeper cross-modal interaction and register-augmented refinement progressively enhance feature discriminability. Our final model adopts $N = 3$ to further stabilize training without overfitting, confirming that the proposed RA-CME layers effectively capture both intra- and inter-modal dependencies.

Table 5 Impact of RA-CME layer count N on MOSI.

N	Acc-2	F1	MAE↓	Corr↑
1	86.88	86.86	0.6026	0.8505
2	86.88	86.81	0.5956	0.8571
3	**87.17**	**87.16**	**0.5810**	**0.8610**

4.3.3 Gating Mechanism Ablation

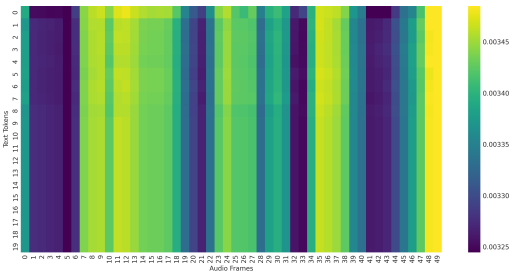
We compare prototype-guided gating against static and no-gate variants (Table 6). The score-guided gate achieves the strongest results (Acc-2 0.8688, F1 0.8681, MAE 0.5878, Corr 0.8607), significantly outperforming the no-gate baseline



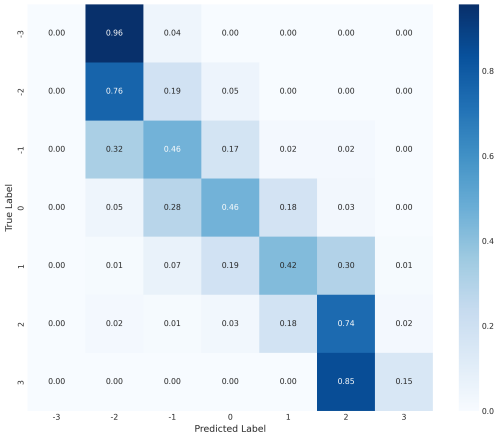
(a) MMML (MOSI)



(b) RegCal-Net (MOSI)



(c) Cross-modal attention



(d) Confusion matrix

Fig. 2 Qualitative analysis on MOSI. (a)–(b) t-SNE visualization of fused features. (c) Cross-modal attention map. (d) Confusion matrix on the test set.

and even the vector-gate variant. This demonstrates that dynamically assessing fused-feature quality via the learned sentiment prototype is crucial for suppressing conflicting signals. In contrast, static or absent gating leads to higher error accumulation, underscoring the necessity of our SCF module for robust bimodal fusion.

4.4 Qualitative Analysis

To visually validate the effectiveness of the Register-Augmented Self-Attention (RASA) mechanism and the Self-Calibrated Fusion (SCF) module, we performed a qualitative analysis of the model's feature space, cross-modal interactions, and decision boundaries.

Table 6 Gating mechanism ablation on MOSI.

Variant	Acc-2	F1	MAE↓	Corr↑
No gate	0.8513	0.8515	0.6061	0.8507
Scalar gate	0.8571	0.8553	0.6010	0.8518
Vector gate	0.8601	0.8587	0.5975	0.8565
Score-guided Gate	**0.8688**	**0.8681**	**0.5878**	**0.8607**

4.4.1 Feature Separability Analysis

We utilize the t-SNE technique to project the fused features extracted by the baseline model and RegCal-Net onto a 2D plane. As shown in Fig. 2(a), the baseline exhibits heavy overlap

between classes. In contrast, Fig. 2(b) shows more separable clusters for RegCal-Net.

4.4.2 Cross-Modal Attention Analysis

Fig. 2(c) reveals the dynamic fusion strategy characterized by distinct vertical attention bands. For example, bright vertical bands indicate globally salient acoustic events attended by many text tokens, whereas darker regions are suppressed as background noise.

4.4.3 Confusion Matrix Analysis

Fig. 2(d) shows strong performance on high-intensity classes, with most errors concentrated between adjacent low-intensity classes, reflecting the intrinsic ambiguity of fine-grained sentiment boundaries.

5 Conclusion

In this work, we proposed RegCal-Net, a bimodal sentiment analysis framework that integrates register-augmented attention and self-calibrated fusion mechanisms. The proposed RASA module improves representation stability by introducing register tokens as global anchors, while the SCF module dynamically evaluates the reliability of fused multimodal features. Extensive experiments on two widely used multimodal sentiment benchmarks demonstrate that the proposed framework consistently improves fine-grained sentiment prediction performance.

Future work will focus on extending the proposed architecture to tri-modal scenarios by incorporating visual cues, as well as exploring lightweight variants for real-time multimodal sentiment analysis applications.

Funding

Funding information is omitted for double-blind review.

References

[1] Dhal, D., Mohanty, J.P., Rout, S.S.: A comprehensive survey on multimodal sentiment analysis. In: 2024 International Conference on Intelligent Computing and Sustainable Innovations in Technology (IC-SIT), pp. 1–6 (2024). IEEE

[2] Lin, J.C.-W., Lv, Q., Yu, D., Srivastava, G., Chen, C.-H.: Optimized scheduling of resource-constraints in projects for smart construction. Information Processing & Management **59**(5), 103003 (2022)

[3] Srivastava, G., Bag, S.: Modern-day marketing concepts based on face recognition and neuro-marketing: A review and future research directions. Benchmarking: An International Journal **31**(2), 410–438 (2024)

[4] Tsai, Y.-H.H., Bai, S., Liang, P.P., Kolter, J.Z., Morency, L.-P., Salakhutdinov, R.: Multimodal transformer for unaligned multimodal language sequences. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL), pp. 6558–6569 (2019)

[5] 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 (ACM MM), pp. 1122–1131 (2020)

[6] Zadeh, A., Chen, M., Poria, S., Cambria, E., Morency, L.-P.: Tensor fusion network for multimodal sentiment analysis. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1103–1114 (2017)

[7] Liu, Z., Shen, Y., Lakshminarasimhan, V.B., Liang, P.P., Zadeh, A., Morency, L.-P.: Efficient low-rank multimodal fusion with modality-specific factors. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL), pp. 2247–2256 (2018)

[8] Yu, W., Xu, H., Yuan, Z., Wu, J.: Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, pp. 10790–10797 (2021)

[9] Huang, C., Chen, J., Huang, Q., Wang, S., Tu, Y., Huang, X.: Atcaf: Attention-based causality-aware fusion network for multimodal sentiment analysis. Information Fusion **114**, 102725 (2025)

[10] Chen, J., Zhong, Y., Huang, Q., Huang, C., Jiang, F., Huang, X., Wang, X.: UCMIB-PNS: Balancing Sufficiency and Necessity With Probabilistic Causality and Cross-Modal Uncertainty in Multimodal Sentiment Analysis. IEEE Transactions on Affective Computing **17**(01), 92–106 (2026)

[11] Huang, C., Lin, Z., Huang, Q., Huang, X., Jiang, F., Chen, J.: H2can: Heterogeneous hypergraph attention network with counterfactual learning for multimodal sentiment analysis. Complex & Intelligent Systems **11**(04) (2025)

[12] Huang, C., Lin, Z., Han, Z., Huang, Q., Jiang, F., Huang, X.: Pamoe-msa: Polarity-aware mixture of experts network for multimodal sentiment analysis. International Journal of Multimedia Information Retrieval **14**(01) (2025)

[13] Chen, J., Huang, Q., Huang, C., Huang, X.: Actual cause-guided adaptive gradient scaling for balanced multimodal sentiment analysis. ACM Trans. Multimedia Comput. Commun. Appl. **21**(6) (2025)

[14] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)

[15] Darcet, T., Oquab, M., Mairal, J., Bojanowski, P.: Vision transformers need registers. arXiv preprint arXiv:2309.16588 (2023)

[16] Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Computation **9**(8), 1735–1780 (1997)

[17] Wang, K., Zeng, X., Yang, J.: Adaptive gating for multimodal fusion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3241–3250 (2021)

[18] Zadeh, A., Zellers, R., Pincus, E., Morency, L.-P.: Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages. IEEE Intelligent Systems **31**(6), 82–88 (2016)

[19] Zadeh, A.B., Liang, P.P., Poria, S., Cambria, E., Morency, L.-P.: Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL), pp. 2236–2246 (2018)

[20] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: RoBERTa: a robustly optimized BERT pretraining approach. arXiv preprint arXiv:1907.11692 (2019)

[21] zhang, H., Wang, Y., Yin, G., Liu, K., Liu, Y., Yu, T.: Learning language-guided adaptive hyper-modality representation for multimodal sentiment analysis. arXiv preprint arXiv:2310.05804 (2023)

[22] Wang, D., Guo, X., Tian, Y., Liu, J., He, L., Luo, X.: TETFN: A text enhanced transformer fusion network for multimodal sentiment analysis. Pattern Recognition **136**, 109259 (2023)

[23] Zadeh, A., Liang, P.P., Mazumder, N., Poria, S., Cambria, E., Morency, L.-P.: Memory fusion network for multi-view sequential learning. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 32 (2018)

[24] Mai, S., Xing, S., He, J., Zeng, Y., Hu, H.: Analyzing unaligned multimodal sequence via graph convolution and graph pooling fusion. arXiv preprint arXiv:2011.13572 (2020)

[25] Rahman, W., Hasan, M.K., Lee, S., Zadeh, A.B., Mao, C., Morency, L.-P., Hoque, E.: Integrating multimodal information in large pretrained transformers. In: Proceedings of

the 58th Annual Meeting of the Association for Computational Linguistics, pp. 2359–2369 (2020)

[26] Yang, D., Huang, S., Kuang, H., Du, Y., Zhang, L.: Disentangled representation learning for multimodal emotion recognition. In: Proceedings of the 30th ACM International Conference on Multimedia, pp. 1642–1651 (2022)

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

[28] Mai, S., Zeng, Y., Zheng, S., Hu, H.: Hybrid contrastive learning of tri-modal representation for multimodal sentiment analysis. IEEE Transactions on Affective Computing **14**(3), 2276–2289 (2022)

[29] Wu, Z., Gong, Z., Koo, J., Hirschberg, J.: Multimodal multi-loss fusion network for sentiment analysis. In: Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pp. 3588–3602 (2024)