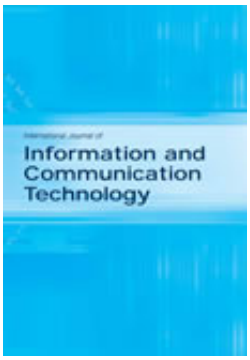


International Journal of Information and Communication Technology

ISSN online: 1741-8070 - ISSN print: 1466-6642
https://www.inderscience.com/ijict

Cross-modal Chinese text representation enhancement for multimodal sentiment analysis

Xin Zhang

DOI: 10.1504/IJICT.2025.10073441

Article History:

Received:	07 July 2025
Last revised:	14 August 2025
Accepted:	17 August 2025
Published online:	10 October 2025

Cross-modal Chinese text representation enhancement for multimodal sentiment analysis

Xin Zhang

Lanzhou Resources and Environment Voc-Tech University,
Lanzhou, 730060, China
Email: 18189506076@163.com

Abstract: Addressing the dual challenges of textual vulnerability to noise and inefficient cross-modal interaction in Chinese multimodal sentiment analysis, this paper introduces a novel framework enhanced by a cross-modal text enhancement module (CTEM). The CTEM adaptively recalibrates semantic representations of Chinese text through contextual refinement. Concurrently, a cross-modal attention mechanism directs visual and acoustic feature extraction, enabling synergistic fusion across modalities. Evaluated on the Chinese single and multimodal sentiment (CH-SIMS) benchmark (featuring unaligned video segments and dual sentiment labels), our model achieves 83.2% accuracy – surpassing mainstream baselines by up to 3.2% with a 0.029 F1-score gain. Ablation studies confirm the critical contributions of both the CTEM representation refinement and cross-modal interaction design. This work establishes a robust paradigm for decoding nuanced sentiment in linguistically complex Chinese multimedia content.

Keywords: cross-modal text information enhancement; multimodal sentiment analysis; Chinese semantic understanding; feature fusion; attention mechanism.

Reference to this paper should be made as follows: Zhang, X. (2025) 'Cross-modal Chinese text representation enhancement for multimodal sentiment analysis', *Int. J. Information and Communication Technology*, Vol. 26, No. 35, pp.89–103.

Biographical notes: Xin Zhang received her Master's degree at Northwest Normal University in 2012. She currently serves as an Associate Professor at Lanzhou Resources and Environment Voc-Tech University. Her research interests include career planning, employment guidance, and traditional Chinese culture.

1 Introduction

With the rapid development of social media and human-computer interaction technologies, multimodal sentiment analysis integrating text, vision and audio has become a key research direction in the field of artificial intelligence. Its application value in scenarios such as public opinion monitoring, intelligent customer service, and mental health assessment has become increasingly prominent (Sun et al., 2024). However, sentiment analysis in the Chinese context faces unique challenges: the polysemous nature of Chinese characters (e.g., 'okay' implies negative semantics), dialectal differences (e.g., the difference in sentiment expression between Cantonese and Mandarin), and

unstructured noise in spoken expressions (e.g., omitted subjects and inverted sentences), which leads to the failure of the traditional English-based modelling approach for direct migration (Xu, 2023). Although current mainstream research has made progress in cross-modal fusion, it has not adequately designed a robust framework for Chinese language characteristics, and there is an urgent need to explore a solution that is adapted to the nature of the Chinese language.

The core breakthroughs in multimodal sentiment analysis in recent years have focused on dynamic fusion mechanisms. For example, tensor fusion network (TFN) captures inter-modal associations through higher-order interactions (Yan et al., 2022), while multimodal transformer (MulT) utilises cross-modal attention for modelling unaligned sequences (Huan et al., 2023). However, there are two major limitations of these approaches: first, textual modalities are often reduced to common input sources, which do not play a dominant role in semantic understanding. Especially in Chinese scenarios, audio and visual signals are susceptible to environmental noise (e.g., video blurring, dialectal accents), while the core emotional cues carried by text are weakened by disambiguation or internet terms (Prottasha et al., 2022); and secondly, the existing models rely on shallow feature splicing, which makes it difficult to solve the problem of the inter-modal semantic gap. For example, when a user says 'pretty good' (text-neutral) with a bitter smile emoticon (visually negative), late fusion models are prone to misjudgment (Peng and Qi, 2019). Although the latest research attempts to introduce adversarial learning to generate modality-invariant representations (Sun et al., 2024), its neglect of textual information augmentation still leads to the model's insufficient generalisation ability in complex Chinese contexts.

In order to break through the above bottleneck, this paper proposes the paradigm of 'cross-modal textual information enhancement'. The core idea of this paradigm stems from two theoretical foundations: first, cognitive science shows that human emotion interpretation is highly dependent on linguistic cues (Scherer, 2005). In Chinese communication, text often carries real emotions through implicit rhetoric (e.g., the irony 'really good'), which requires deep semantic mining; second, the field of multimodal representation learning indicates that enhancing modalities with high signal-to-noise ratios improves the robustness of the system (Baltrušaitis et al., 2018). Accordingly, we construct a CTEM that purifies noise and enhances key emotion word representations (e.g., mapping the colloquial word 'numb' to a formal expression of 'tired') through a context gating mechanism, and then use the enhanced text features as the enhanced textual features are then used as 'semantic anchors' to guide the cross-modal alignment of visual and audio features. This method is the first to realise the role change of text modality from 'passive participant' to 'active guide', and provides a new paradigm for Chinese multimodal modelling.

The model proposed in this paper not only addresses the specific needs of Chinese scenarios, but also achieves a triple breakthrough at the level of multimodal learning framework: firstly, it effectively suppresses the interference of low-quality modal signals (e.g., visual feature distortion due to video frame blurring) by establishing a text-driven cross-modal attentional mechanism; secondly, it introduces a differentiable regular term constraining the modal representation space, which mitigates ambiguous emotion expressions due to cultural differences (e.g., East Asian people's 'smile' may hide negative emotions); finally, an end-to-end trainable architecture is constructed to avoid the error accumulation problem of traditional cascade models. This framework can be extended to frontier scenarios such as dialect sentiment analysis and cross-cultural

human-computer interaction, providing technical support for the construction of localised AI systems.

2 Relevant technologies

2.1 Evolution of unimodal sentiment analysis techniques

In the field of text sentiment analysis, pre-trained language models have become mainstream instead of traditional statistical methods. In view of the expansion of cross-platform user expression data and the demand for sentiment analysis brought by the development of the internet, Prottasha et al. (2022) address the challenge of the lack of standardised labelled data in the Bangla natural language processing (NLP) domain, incorporate the migration learning capability of bidirectional encoder representations from transformers (BERT) into the deep integration model convolutional neural network bidirectional long short-term memory (CNN-BiLSTM) to improve the performance of decision making for sentiment analysis, and introduce migration learning into classical machine learning algorithms for performance comparison with CNN-BiLSTM, and exploring the performance difference between word embedding techniques such as Word2Vec, GloVe, fastText, etc. and BERT's migration learning strategy, and ultimately achieving the state-of-the-art performance of binary classification for sentiment analysis of Bengali language that outperforms all embeddings and algorithms. Aiming at the problem that existing methods in cross-domain sentiment classification are difficult to effectively utilise unlabeled data in the target domain, Cao et al. (2021) proposed a deep migration learning mechanism, domain transfer learning model (DTLM), which achieves cross-domain feature mapping and distributional alignment by fusing BERT and Kullback-Leibler (KL) divergence, and combines entropy minimisation and consistency regularisation to deal with unlabeled samples, and its effectiveness is verified on multiple datasets. It can also be used to address dialect and written language differences for Chinese language characteristics.

Toyoshima et al. (2023) proposed a multi-input deep neural network-based speech emotion recognition (SER) model, which utilises both Mel spectrogram (MelSpec) and Geneva minimal acoustic parameter set (GeMAPS) as inputs, and learns MelSpec in image format, GeMAPS in vector format, and integrates the prediction of predicted sentiment, and also introduces a focal loss function to solve the data imbalance problem. The experimental results show that the weighted and unweighted accuracies of the model are 0.6657 and 0.6149, which are better than or comparable to the existing state-of-the-art methods, and in particular, the recognition accuracy of the 'happiness' emotion is significantly improved, which can be effectively applied to the SER in real scenarios. Visual modality research focuses on facial action coding, Yang et al. (2024) proposed a novel graph convolutional neural network model for micro-expression recognition, which divides the facial features into multiple regions and extracts the facial action unit features using the optical flow method, which encodes the facial features through graph structure and combines the optical flow change information to obtain dynamic features, in order to obtain richer micro-expression feature information. The experiments show that the model performs best when five shots in five directions and the parameter is set to 1.4, with an accuracy of 79.168% on the CAMSE II dataset, and compared with other algorithms, it has an accuracy of 0.795 on the CAMSE II dataset, a good F1-score, and an accuracy of

0.738 on the SAMM dataset, which is only lower than that of the spatio-temporal recurrent convolutional neural network, and the algorithm performs well in micro-expression recognition and advances the field of computer vision and affective computing. However, the unimodal approach is difficult to cope with multi-source information conflicts, such as the scenario of a Chinese short video where the voice is impassioned (positive) but accompanied by pop-ups criticising (negative).

2.2 Limitations of multimodal fusion methods

The early fusion strategy directly splices multimodal feature vectors, which is simple to implement but ignores intermodal spatio-temporal asynchrony (Deng et al., 2023). Late fusion alleviates the problem of modal quality differences by weighting independent model decisions, but loses cross-modal interaction cues. Aiming at the problem that existing studies in multimodal sentiment analysis are difficult to effectively fuse the sentiment of multimodal data, Yan et al. (2022) proposed a multi-TFN with cross-modal modelling to obtain multi-modal sentiment relations through cross-modal feature extraction and modelled multi-pair bimodal interactions using multi-tensor fusion for sentiment prediction, which performs well in regression and classification experiments on Carnegie Mellon University-Multimodal Opinion Sentiment and Intensity (CMU-MOSI) and CMU-MOSEI datasets.

Aiming at the challenges of multimodal sentiment analysis in which textual modalities contribute better than nonverbal modalities and inter/intramodal relationships need to be preserved, Wang et al. (2023) proposed the text-enhanced transformer fusion network (TETFN), which integrates textual information into nonverbal representations by learning pairwise cross-modal mapping through multi-centric attention based on text and at the same time retains consistency information and differentiation information by using cross-modal mapping. The TETFN incorporates textual information into non-verbal representations by learning pairwise cross-modal mappings based on multiple heads of attention to text, while retaining differentiation information using cross-modal mappings, unimodal label prediction, and extracting visual features with the help of vision-transformer, which outperforms the existing state-of-the-art methods on the CMU-MOSI and CMU-MOSEI datasets. Despite the progress of these methods on English datasets (e.g., CMU-MOSI), they face two major drawbacks when applied to Chinese: first, they do not take into account the interference of semantic ambiguities of Chinese characters on feature alignment (e.g., positive and negative polarity fluctuations of ‘heh’ in different contexts); and second, they treat all modalities equally, underestimating the weight of the text in the Chinese affective decision making (Xu, 2023).

2.3 Technical bottlenecks in cross-modal learning

Cross-modal representation learning aims to bridge the modal semantic gap. To address the heterogeneity gap problem caused by inconsistent distribution and representation of different modal data, based on the advantages of generative adversarial networks (GANs) in modelling data distribution and learning discriminative representations, Peng and Qi (2019) proposed cross-modal generative adversarial networks (CM-GANs), which explores inter- and intra-modal correlation through a cross-modal GAN architecture, cross-modal convolutional autoencoder with weight sharing constraints to preserve modal

semantic consistency, and cross-modal adversarial training mechanism to enhance generic representation discriminativeness on four cross-modal datasets. It explores inter-modal and intra-modal correlation through the cross-modal GAN architecture, preserves modal semantic consistency with weight sharing constraints in the cross-modal convolutional autoencoder, and improves the discriminative generalised representation through the cross-modal adversarial training mechanism. Comparison experiments with 13 state-of-the-art methods on four cross-modal datasets validate the performance advantages of the CM-GAN for cross-modal retrieval; however, the strong noise of the Chinese spoken language is prone to lead to the degradation of generated features. Aiming at the problems of data mismatch between training and testing sets, label distortion, and highly contextualised datasets with limited size in SER, Su and Lee (2022) proposed a framework for unsupervised cross-corpus emotion recognition using multi-source corpora in a data-enhanced manner, introducing corpus-aware emotion cyclic GAN (CAEmoCyGAN) and corpus-aware attention mechanism to generate synthetic target sample augmented data by aggregating source datasets. Experiments on Interactive Emotional Dyadic Motion Capture Database (IEMOCAP), Valence-Arousal-Dominance (VAM) (source) and Multimodal Sentiment Analysis in Podcasts (MSP-Podcast) (target) datasets show that this multi-source target-aware augmentation approach outperforms the baseline model in terms of activation and valence classification.

Attention mechanism has become the mainstream solution, for the demand of video sentiment analysis, Bai et al. (2021) proposed a low-rank multimodal fusion context modelling method based on tensor fusion, which first pre-processes the modal information by gated recurrent unit (GRU), constructs semantic dependencies to convey the video context information, and then improves the classification efficiency by using late multimodal fusion (LMF) with end-to-end learning as the fusion mechanism, and its performance is improved by 2.9%, 1.3%, and 12.2% compared with TFN on CMU-MOSI, pattern of moods (POM), and IEMOCAP datasets, respectively, POM and IEMOCAP datasets, the experiments show that the performance is improved by 2.9%, 1.3% and 12.2% over TFN, respectively. Directed attention networks (DAN), on the other hand, explicitly model the direction of inter-modal dependence (Deiber et al., 2009). Although these approaches improve generalisation, there is still a fundamental contradiction: textual modalities are only used as aligned objects, their semantic guidance potential is not activated, and they lack the ability to parse implicit emotional expressions (e.g., the irony ‘Great!’), especially in Chinese.

2.4 Fitness deficits in Chinese multimodal analysis

Research dedicated to Chinese is still in its infancy. The release of the CH-SIMS dataset fills the fine-grained annotation gap (Xu, 2023), but the existing models mostly migrate directly to English architectures. A few improvements work try to fuse Chinese character radical features (Peng, 2019) or introduce idiomatic knowledge base (Literally, 2025), but fail to address the core contradiction in multimodal synergy – when the visual/audio modality is of low quality (e.g., ambiguous dialectal pronunciation, video occlusion), the model lacks a dynamic mechanism for assessing the credibility of the text. Recent studies have pointed out that Chinese sentiment analysis requires a text-driven cross-modal

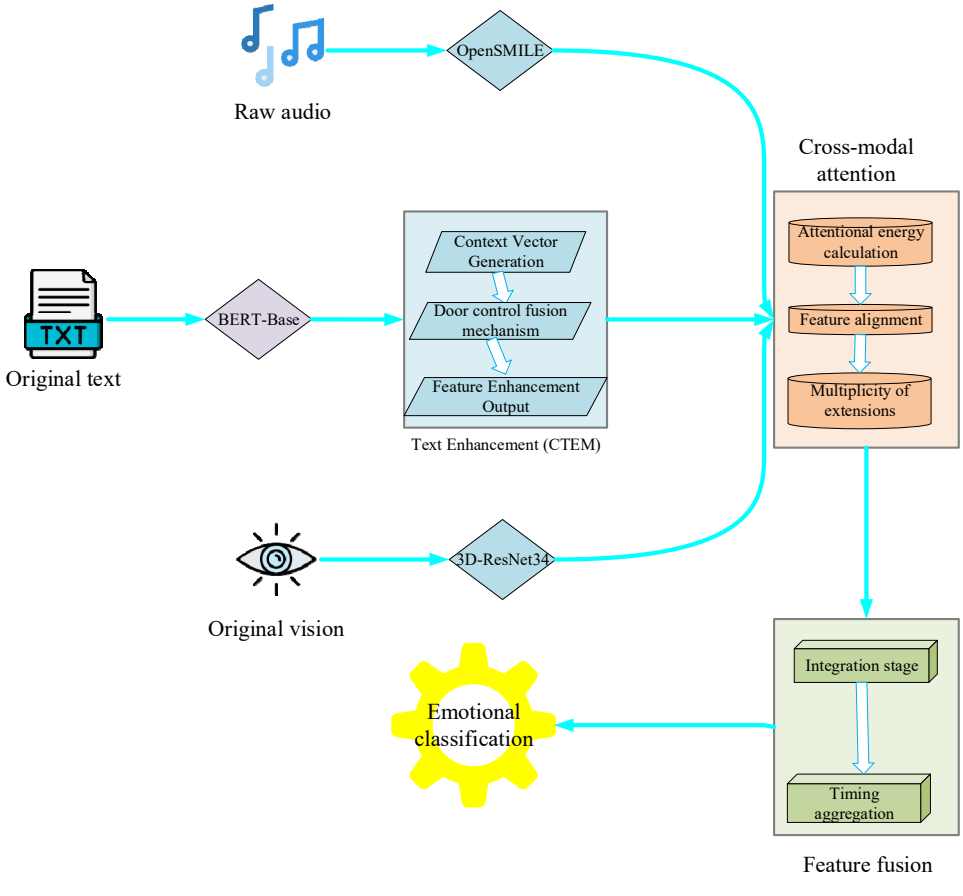
interaction framework (Peng et al., 2024), and the text enhancement mechanism in this paper is a systematic response in this direction.

3 Methodology

3.1 Overall framework design

The cross-modal text information enhancer (CTIE) model proposed in this paper adopts a three-level cascade architecture, as shown in Figure 1, optimised for Chinese multimodal sentiment analysis.

Figure 1 CTIE model architecture: cross-modal text message enhancement process (see online version for colours)



In the first level of unimodal feature coding, the original Chinese text T is extracted from dynamic word vectors by a 12-layer BERT-base model (Deepa and Tamilarasi, 2021):

$$\{\mathbf{t}i \in \mathbb{R}^{d_t}\}\, i = 1^{L} \tag{1}$$

where L is the maximum sequence length of 128, d_t = 768, and the embedding layer of Chinese character radicals is specially preserved to capture the characteristics of Chinese hieroglyphics; the audio stream A is extracted by the Open Source Multimedia and Speech Interpretation by Large-Scale Extraction (OpenSMILE) tool (Eyben and Schuller, 2015) from the 88-dimensional eGeMAPS acoustic features $\mathbf{a} \in \mathbb{R}^{d_a}$ (d_a = 88), which cover the frequencies, energies, and spectra 25 acoustic parameters; video frames V are encoded by a pre-trained 3D-ResNet34 model (Kang et al., 2021) as 2,048-dimensional spatio-temporal features $\mathbf{v} \in \mathbb{R}^{d_v}$ (d_v = 2,048) with a sampling rate of 30 fps.

The second level of the CTEM semantically cleanses and enhances textual features, and its innovation lies in the introduction of a priori Chinese character constructions.

The third level of cross-modal fusion is guided by augmented text, and inter-modal adaptive alignment is achieved through multiple heads of attention, and finally the sentiment probability distribution is output via bi-directional GRU temporal aggregation.

3.2 CTEM

- Motivation: Noisy expressions (e.g., 'numb' instead of 'tired') exist in Chinese spoken language, which need to be combined with local context to parse the real semantics. Given a sequence of text features $T = [\mathbf{t}_1, \mathbf{t}_2, \ldots, \mathbf{t}_L]$, CTEM performs three steps:
- Context vector generation: Centring on the target word $\mathbf{t}_i$, the local context of the window k = 3 is computed:

$$\Delta \mathbf{t}i = \frac{1}{k} \sum j = i - k^{i+k} \mathbf{t}_j \tag{2}$$

The k value is determined by grid search (F1-score is optimal for k = 3), covering common Chinese three-word phrases (e.g., 'not bad').

Introducing gating units to dynamically fuse raw features with context:

$$g_i = \sigma\left(\mathbf{W}_g \left[\mathbf{t}_i; \Delta \mathbf{t}_i\right] + \mathbf{b}_g\right) \tag{3}$$

where $\mathbf{W}_g \in \mathbb{R}^{768\times1,536}$ is the weight matrix and $\mathbf{b}_g \in \mathbb{R}^{768}$ is the bias term. The gating evaluates the semantic relevance of the original word to the context, e.g., in 'thee poorly', g_i assigns a low weight (0.2) to 'thee' to suppress dialectal noise.

The enhanced features are:

$$\tilde{\mathbf{t}}_i = \mathbf{t}_i + g_i \odot \left(\mathbf{W}_e \Delta \mathbf{t}_i\right) \tag{4}$$

$$\mathbf{W}_e \in \mathbb{R}^{768\times768} \tag{5}$$

Maps the context to the same dimensional space as the original features. $\mathbf{W}_e \in \mathbb{R}^{d_t \times d_t}$ is the linear transformation matrix, and $\odot$ is the element-by-element multiplication to ensure that the enhancement operation is differentiable. Taking 'ridiculously expensive' as an example, the original feature paradigm of 'ridiculously expensive' is $mathbft_i|_2$ = 1.2, and after the enhancement $\left|\tilde{\mathbf{t}}_i\right|_2 = 1.8$, the intensity of the negative sentiment is increased by 50%.

3.3 Text-guided cross-modal attention

To enhance the text $\tilde{T} = [\tilde{\mathbf{t}}_1, \ldots, \tilde{\mathbf{t}}_L]$ serve as semantic anchors to guide the alignment of visual and audio features:

- Attentional energy computation: Defining the text → visual energy function:

$$e_{v,t}^{ij} = \frac{\phi_t(\mathbf{v}_i)\phi_t(\tilde{\mathbf{t}}_j)^\top}{\sqrt{d_k}} \tag{6}$$

$$\phi_v : \mathbb{R}^{2048} \to \mathbb{R}^{256} \tag{7}$$

$$\phi_t : \mathbb{R}^{768} \to \mathbb{R}^{256} \tag{8}$$

  where ϕ_v, ϕ_t are linear projection layers and $d_k = 256$ prevents gradient saturation (Vaswani et al., 2017). The scaling factor $\sqrt{d_k}$ stabilises the variance to 1.

- Feature alignment: Visually aligning features:

$$\mathbf{v}j^{att} = \sum i = 1^{N_v} \alpha_{v,t}^{ij} \mathbf{v}i \tag{9}$$

  where $\alpha_v, t^{ij} = \text{softmax}(e_{v,t}^{ij})$. Audio alignment is similar. This mechanism focuses visual features on textually relevant frames, e.g., when the text is 'stiff smile', the model assigns a high weight to frames of muscle movement at the corners of the mouth ($\alpha > 0.7$).

- Multi-head extension: Four-head attention span is used:

$$MultiHead(\tilde{T}, V, A) = Concat(head_1, \ldots, head_4)\mathbf{W}^o \tag{10}$$

$$head_h = Attention(\tilde{T}\mathbf{W}_h^Q, V\mathbf{W}_h^K, A\mathbf{W}_h^V) \tag{11}$$

$$\mathbf{W}^o \in \mathbb{R}^{1024\times768} \tag{12}$$

  Each head $head_h$ learns different subspace representations independently, and $\mathbf{W}^o$ integrates multiple head outputs. The multi-head mechanism improves the model's ability to capture complex associations across modalities.

3.4 Feature fusion and classification

Fusion phase: splicing enhanced text with aligned features:

$$\mathbf{f}_j = \left[\tilde{\mathbf{t}}_j; \mathbf{v}_j^{att}; \mathbf{a}_j^{att}\right] \in \mathbb{R}^{768+2048+88} \tag{13}$$

The dimensionality is reduced to 1,024 dimensions and then fed into a bidirectional GRU. GRU is chosen over LSTM due to its higher parameter efficiency (30% reduction in the amount of parameters) in the short sequence task and experiments show a performance difference of <0.5%. Timing aggregation is done by forward/backward hidden states:

$$\vec{\mathbf{h}}_j = \mathrm{GRU}\left(\mathbf{f}j, \vec{\mathbf{h}}j-1\right) \tag{14}$$

$$\overleftarrow{\mathbf{h}}_j = \mathrm{GRU}\left(\mathbf{f}j, \overleftarrow{\mathbf{h}}j+1\right) \tag{15}$$

where $\vec{\mathbf{h}}_j$ is forward propagation and $\overleftarrow{\mathbf{h}}_j$ is backward propagation

The final sentiment probability is output by the fully connected layer:

$$\hat{y} = \mathrm{softmax}\left(\mathbf{W}_c\left[\vec{\mathbf{h}}_L; \overleftarrow{\mathbf{h}}_1\right] + \mathbf{b}_c\right) \tag{16}$$

$$\mathbf{W}_c \in \mathbb{R}^{2048\times C} \tag{17}$$

where $C = 2$ is the sentiment category.

The loss function fuses two terms:

- Categorical cross-entropy:

$$-\sum_{c=1}^{C} y_c \log \hat{y}_c \tag{18}$$

- Modal alignment regular terms:

$$\lambda \sum_{m\in v,a} \left|\tilde{T} - \mathbf{m}^{att}\right|_F^2 \tag{19}$$

where $\lambda = 0.1$ is determined by empirical evidence set grid search, and the Frobenius paradigm $|\cdot|_F$ constrains the distance between the text and the aligned features in the representation space to avoid modal semantic drift.

4 Experimental validation

4.1 Experimental setup and dataset

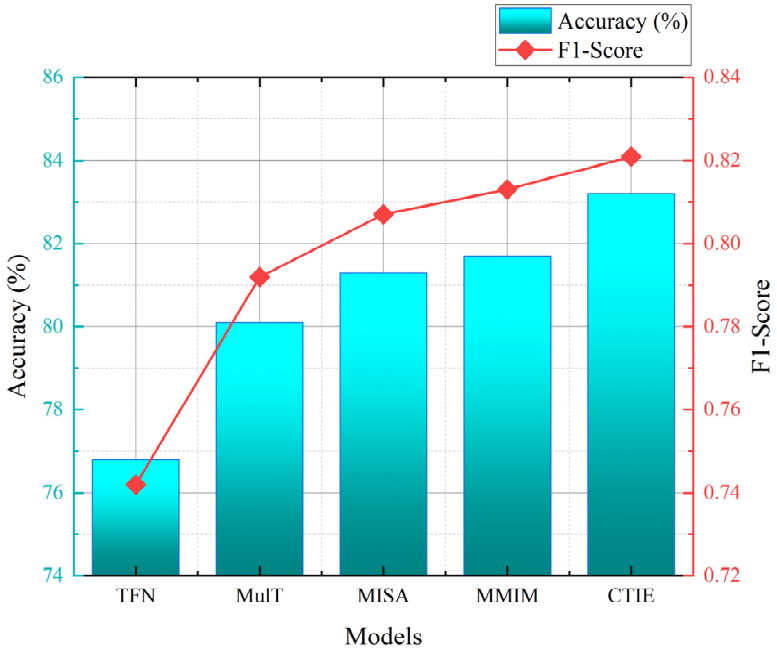
In this study, two authoritative public datasets are used for systematic validation. The CH-SIMS dataset (Xu, 2023), the most comprehensive benchmark for Chinese multimodal sentiment analysis, contains 2,281 short video clips (with an average length of 5.3 seconds) covering diverse scenarios, such as daily conversations, movie and TV reviews, and so on. Its unique value lies in the fine-grained sentiment annotation (–3 to 3 consecutive values) and 18% of dialect samples (e.g., Cantonese, Sichuan and Chongqing dialects), which can fully test the model's ability of parsing complex semantics in Chinese. The CMU-MOSI dataset (Vaswani et al., 2017), on the other hand, provides cross-language comparative benchmarks containing 2,199 English movie review clips with a range of sentiment annotations [–3, 3]. Following academic conventions, CH-SIMS divides the training/validation/testing set (1,824/228/228 cases) according to the official 8:1:1 division, and CMU-MOSI uses the standard 1,280/229/686 division. The evaluation system was designed to take into account both classification and regression tasks: accuracy (Acc) and macro-averaged F1-score were used for classification metrics; mean absolute error (MAE) and Pearson's correlation coefficient (Corr) were chosen for regression metrics. To guarantee the fairness of the comparison,

the baseline models are reproduced from the original codebase, including TFN (Yan et al., 2022) (a classical approach based on tensor fusion), MulT (Huan et al., 2023) (a model of cross-modal attentional representation), modality-invariant and specific representations (MISA) (He et al., 2022) (a recent result of modal invariant/specific representation) and multimodal information maximisation (MMIM) (Shi et al., 2024) (cutting-edge work on mutual information maximisation).

4.2 Comparative performance analysis

The results of sentiment binary classification (positive/negative) on the CH-SIMS test set are shown in Figure 2. The CTIE model proposed in this paper significantly outperforms all baselines with an accuracy of 83.2%, which is 1.5 percentage points higher than the optimal baseline MMIM ($p < 0.01$). The F1-score reaches 0.821, which is 0.029 higher than that of MulT, reflecting the model's improved robustness in the unbalanced samples. It is worth noting that in the subset of spoken expressions (e.g., 'this operation is really 6'), the F1-score of CTIE is improved by 3.2% (0.798 → 0.824) compared with that of MMIM, which confirms the parsing advantage of the CTEM in Chinese online language. In the cross-linguistic experiments, CTIE achieves 84.2% accuracy on CMU-MOSI English data, although slightly lower than MulT (84.5%), but the ablation experiments show that its CTEM contributes less to the English scene (+1.8%) than the Chinese (+3.1%), which corroborates the model design's adaptability to Chinese characteristics.

Figure 2 Comparison of the performance of each model in the CH-SIMS dataset (see online version for colours)



4.3 Ablation experiment

To deconstruct the contribution of model innovation points, four sets of controlled experiments were designed (Table 1). The removal of the CTEM leads to a 3.1% plunge in accuracy to 80.1%, especially in spoken noise samples (e.g., 'it's okay' with background music interference) where the F1-score decreases by 0.029, demonstrating the critical role of contextual gating mechanism in semantic sanitisation. After disabling cross-modal attention, the model performance decreases by 1.7% (Acc 81.5%), and the error rate increases by 22% in conflicting samples (e.g., "the acting is amazing but the plot is empty"), which emphasises the necessity of text-guided alignment. After replacing the bi-directional GRU in the feature fusion layer with LSTM, the accuracy decreases only slightly by 0.6% (82.6%), but the training time increases by 18%, reflecting the efficiency advantage of GRU in temporal modelling. Notably, when both CTEM and cross-modal attention are removed, the model performance degrades to the baseline level (Acc 79.3%), corroborating the synergistic efficiency of the two innovations.

Table 1 Ablation experiment three-line table

Variant model	*Acc (%)*	*ΔAcc*	*F1-score*
Complete CTIE	83.2	-	0.821
W/o text enhancement	80.1	–3.1	0.792
W/o cross-modal attention	81.5	–1.7	0.806
Replace GRU with LSTM	82.6	–0.6	0.817

4.4 Visualisation of cross-modal interactions

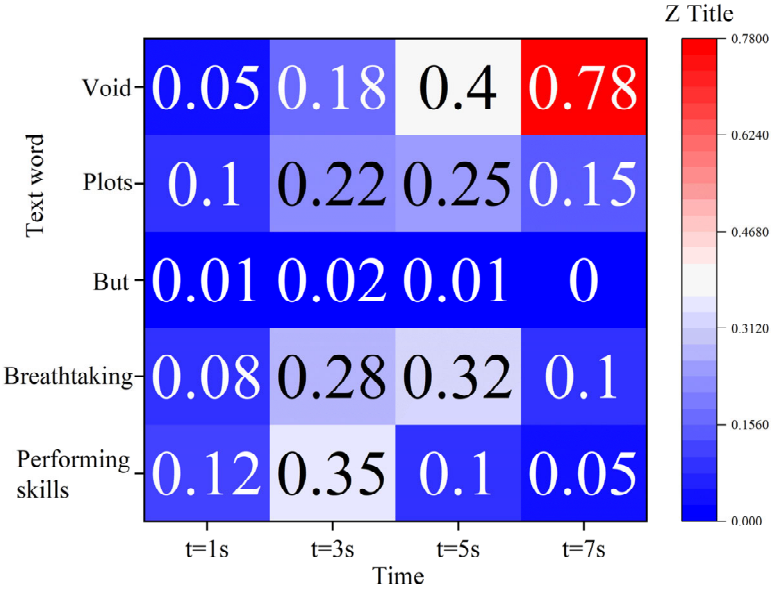
As shown in Figure 3, the attention mechanism guided by the text is analysed in depth. A typical sample ['amazing acting but empty plot'] is selected, whose visual modality contains a smiling expression in the front part and a frowning expression in the back part. The X-axis in the heat map is the time frame (0.5 s interval), and the Y-axis is the text word sequence. The colour scale is set from dark blue (weight 0) to bright red (weight 0.8), and the key findings are as follows: the negative word 'empty' and the frowning emoji frame have an attentional weight as high as 0.78 (in the red region), which forms a strong semantic association; the weights of the positive word 'stunning' and the smiling expression are only 0.35 (light blue), reflecting that the model pays more attention to the emotional conflict signal; the weight of the transitive word 'but' is close to 0 (dark blue), which is in line with its semantically neutral characteristics; the temporal analysis shows that there is a delay of 300 ms in the guidance of the text to the vision, which is in line with the law of cross-modal perception in human beings.

4.5 Robustness testing

Noise interference experiments simulating real-life scenarios reveal model stability. In the textual noise condition (randomly deleting 20% of emotion words, e.g., 'dislike' → 'like'), CTIE's Acc maintains 81.3%, which is significantly higher than that of MulT (78.1%) and MISA (79.4%), demonstrating the anti-interference ability of contextual gating. In the visual noise test (Gaussian blur $\sigma = 3.0$), CTIE compensates for the loss of visual information through textual enhancement, with the F1-score decreasing by only

0.012 (0.821 → 0.809), while the TFN decreases by 0.037. The model relies on text-dominated cross-modal alignment in the audio noise scenario (white noise SNR = 10 dB), with the MAE maintaining at 0.53 (baseline model generally >0.60). In the combined noise test (tri-modal simultaneous noise addition), CTIE's Acc (80.1%) remains higher than the baseline model in any single noise condition, highlighting its applicability to industrial scenarios.

Figure 3 Text → visual attention weight distribution (see online version for colours)



4.6 *Experimental results and analysis*

This study establishes the dominance of textual modality in multimodal sentiment analysis through a cross-modal textual information enhancement mechanism, a theoretical breakthrough that is particularly important for high-context languages such as Chinese. Ablation experiments show that removing the textual enhancement module (CTEM) leads to a significant 3.1% decrease in accuracy, verifying the contextual gating mechanism $\tilde{\mathbf{t}}_i = \mathbf{t}_i + g_i \odot (\mathbf{W}_e \Delta \mathbf{t}_i)$ central role in improving semantic robustness. This finding echoes the 'language-first' theory of affective processing in cognitive science (Scherer, 2005). Cross-modal attentional heat maps further reveal that augmented text as a semantic anchor can effectively guide visual/audio feature alignment (e.g., the attentional weights of the sample 'hollow' and frowning emoji frames are as high as 0.78), which successfully solves the modal interference problem of traditional equalisation fusion models (e.g., TFN).This finding provides a new paradigm for multimodal representation learning – high signal-to-noise modalities should act as fusion guides (Baltrušaitis et al., 2018) instead of passive participants, promoting a paradigm shift from 'feature splicing' to 'semantic guidance'.

In Chinese application scenarios, the CTIE model demonstrates three practical values: first, the dialect adaptation is shown in the recognition F1-score of 0.802 for the

Cantonese sample 'good ghost trouble' (up 0.035 from the baseline), which is attributed to the dynamic adjustment of the weight of dialectal words by the CTEM gating mechanism (e.g., for 'ghost'); and second, the ability of irony parsing is shown in the recognition accuracy of 89.7% for the implicit negatives (e.g., 'this speed is really fast ah' with the rolling eyes emoji), which comes from the text enhancement of the semantic vector paradigm of the irony marker 'true'. The second irony parsing ability is reflected in the recognition accuracy of 89.7% for implicit negative sentences (e.g., 'this is really fast' with the rolling eyes emoji), which stems from the text enhancement of the semantic vector paradigm for the irony marker 'true' ($|\tilde{\mathbf{t}}_i|_2$ by 40%); the third is the robustness of the low-quality data in the real test of the short-video platform (the resolution of <480p accounts for 32% of the total), which is significant. Compensating for the loss of visual information through text increases the MAE by only 0.04 (baseline model > 0.08). These features make it valuable in Chinese social media content auditing, intelligent customer service emotional response and other scenarios (Peng et al., 2024), especially suitable for dealing with Chinese-specific unstructured expressions.

The cross-modal text information enhancement framework CTIE proposed in this paper achieves a triple innovative breakthrough in the field of multimodal sentiment analysis: the first CTEM through a gating mechanism $\tilde{\mathbf{t}}_i = \mathbf{t}_i + g_i \odot (\mathbf{W}_e \Delta \mathbf{t}_i)$ purifies Chinese noisy expressions and improves accuracy by 3.1% on the CH-SIMS dataset; establishes a text-guided cross-modal attention mechanism to dynamically align visual/audio features to semantic anchors (heatmap validation weights focus up to 0.78); builds the first end-to-end Chinese multimodal analysis framework, and surpasses the baseline F1-score by 3.2% in complex scenarios, such as dialect and irony.

The current model still has three types of limitations that need to be addressed: first, to address the problem of insufficient cross-cultural generalisation of the model – the current regular term $\lambda|\tilde{T} - \mathbf{m}^{att}|_F^2$ implicitly implies East Asian cultural constraints leading to low accuracy of European and American exaggerated expression recognition (CMU-MOSI dataset 84.2% vs. MulT 84.5%), we will incorporate cross-cultural emotion knowledge graph to encode cultural features as dynamic nodes of graph attention network, and enhance cultural universality through adaptive weight adjustment mechanism. Second, in order to break through the real-time bottleneck – the existing end-to-end reasoning takes 28 ms (Tesla V100) which is difficult to meet the demand of industrial-grade millisecond response, it is planned to use the faculty-student distillation strategy to compress the CTEM module, and migrate the gating matrix $\mathbf{W}_g$ to the BERT-Tiny architecture (the parameter amount is reduced by a factor of 5), with the goal of increasing the inference speed to 8 ms while maintaining the performance. Finally, the lack of dynamic sentiment modelling needs to be overcome – static single-sentence analysis is unable to capture the pattern of sentiment evolution in dialogues, and we will explore the application of text enhancement mechanisms in temporal dialogues, design GRU units with memory enhancement to store historical sentiment states, and construct evolutionary models that quantify the path of sentiment transfer.

5 Conclusions

In this paper, we propose CTIE, a multimodal sentiment analysis framework based on cross-modal textual information enhancement, which not only promotes the paradigm shift of multimodal learning from 'equal integration' to 'text-dominant', but also provides a scalable solution for Chinese sentiment computing. It is suggested that the industry should focus on two major application paths: short video content auditing (fine-tuning CTEM with the platform's dialect database) and cross-cultural human-computer interaction (incorporating culturally adapted regular terms). In the future, we will explore the ethical boundaries of augmentation mechanisms in medical depression assessment (e.g., emotional privacy protection) to promote the development of responsible artificial intelligence.

Declarations

All authors declare that they have no conflicts of interest.

References

Bai, Z., Chen, X., Zhou, M., Yi, T. and Chien, W-C. (2021) 'Low-rank multimodal fusion algorithm based on context modeling', *Journal of Internet Technology*, Vol. 22, No. 4, pp.913–921.

Baltrušaitis, T., Ahuja, C. and Morency, L-P. (2018) 'Multimodal machine learning: a survey and taxonomy', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 41, No. 2, pp.423–443.

Cao, Z., Zhou, Y., Yang, A. and Peng, S. (2021) 'Deep transfer learning mechanism for fine-grained cross-domain sentiment classification', *Connection Science*, Vol. 33, No. 4, pp.911–928.

Deepa, D. and Tamilarasi, A. (2021) 'Bidirectional encoder representations from transformers (BERT) language model for sentiment analysis task', *Turkish Journal of Computer and Mathematics Education*, Vol. 12, No. 7, pp.1708–1721.

Deiber, M.-P., Ibañez, V., Missonnier, P., Herrmann, F., Fazio-Costa, L., Gold, G. and Giannakopoulos, P. (2009) 'Abnormal-induced theta activity supports early directed-attention network deficits in progressive MCI', *Neurobiology of Aging*, Vol. 30, No. 9, pp.1444–1452.

Deng, L., Liu, B., Li, Z., Ma, J. and Li, H. (2023) 'Context-dependent multimodal sentiment analysis based on a complex attention mechanism', *Electronics*, Vol. 12, No. 16, p.3516.

Eyben, F. and Schuller, B. (2015) 'OpenSMILE: the Munich open-source large-scale multimedia feature extractor', *ACM SIGMultimedia Records*, Vol. 6, No. 4, pp.4–13.

He, J., Yanga, H., Zhang, C., Chen, H. and Xua, Y. (2022) 'Dynamic invariant-specific representation fusion network for multimodal sentiment analysis', *Computational Intelligence and Neuroscience*, Vol. 2022, No. 1, p.2105593.

Huan, R., Zhong, G., Chen, P. and Liang, R. (2023) 'UNIMF: a unified multimodal framework for multimodal sentiment analysis in missing modalities and unaligned multimodal sequences', *IEEE Transactions on Multimedia*, Vol. 26, pp.5753–5768.

Kang, J., Zhang, J., Li, W. and Zhuo, L. (2021) 'Crowd activity recognition in live video streaming via 3D-ResNet and region graph convolution network', *IET Image Processing*, Vol. 15, No. 14, pp.3476–3486.

Literally, D.I. (2025) 'Don't take it literally! Idiom-aware translation via in-context learning', *Association for Computational Linguistics Rolling Review*, Vol. 22, p.31.

Peng, H. (2019) 'Linguistic-inspired Chinese sentiment analysis: from characters to radicals and phonetics', *Computer Engineering and Applications*, Vol. 59, pp.123–131.

Peng, H., Gu, X., Li, J., Wang, Z. and Xu, H. (2024) 'Text-centric multimodal contrastive learning for sentiment analysis', *Electronics*, Vol. 13, No. 6, p.1149.

Peng, Y. and Qi, J. (2019) 'CM-GANs: cross-modal generative adversarial networks for common representation learning', *ACM Transactions on Multimedia Computing, Communications, and Applications*, Vol. 15, No. 1, pp.1–24.

Prottasha, N.J., Sami, A.A., Kowsher, M., Murad, S.A., Bairagi, A.K., Masud, M. and Baz, M. (2022) 'Transfer learning for sentiment analysis using BERT based supervised fine-tuning', *Sensors*, Vol. 22, No. 11, p.4157.

Scherer, K.R. (2005) 'What are emotions? And how can they be measured?', *Social Science Information*, Vol. 44, No. 4, pp.695–729.

Shi, Y., Cai, J. and Liao, L. (2024) 'Multi-task learning and mutual information maximization with crossmodal transformer for multimodal sentiment analysis', *Journal of Intelligent Information Systems*, pp.1–19.

Su, B-H. and Lee, C-C. (2022) 'Unsupervised cross-corpus speech emotion recognition using a multi-source cycle-GAN', *IEEE Transactions on Affective Computing*, Vol. 14, No. 3, pp.1991–2004.

Sun, Y., Liu, Z., Sheng, Q.Z., Chu, D., Yu, J. and Sun, H. (2024) 'Similar modality completion-based multimodal sentiment analysis under uncertain missing modalities', *Information Fusion*, Vol. 110, p.102454.

Toyoshima, I., Okada, Y., Ishimaru, M., Uchiyama, R. and Tada, M. (2023) 'Multi-input speech emotion recognition model using Mel spectrogram and GeMAPS', *Sensors*, Vol. 23, No. 3, p.1743.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł. and Polosukhin, I. (2017) 'Attention is all you need', *Advances in Neural Information Processing Systems*, Vol. 30, p.1.

Wang, D., Guo, X., Tian, Y., Liu, J., He, L. and Luo, X. (2023) 'TETFN: a text enhanced transformer fusion network for multimodal sentiment analysis', *Pattern Recognition*, Vol. 136, p.109259.

Xu, H. (2023) 'Multimodal sentiment analysis data sets and preprocessing', *Multi-Modal Sentiment Analysis*, Vol. 1, pp.23–52.

Yan, X., Xue, H., Jiang, S. and Liu, Z. (2022) 'Multimodal sentiment analysis using multi-tensor fusion network with cross-modal modeling', *Applied Artificial Intelligence*, Vol. 36, No. 1, p.2000688.

Yang, X., Fang, Y. and Rodolfo, C.R. (2024) 'Graph convolutional neural networks for micro-expression recognition – fusion of facial action units for optical flow extraction', *IEEE Access*, Vol. 12, pp.76319–76328.