

TGSE-DFN: Text-guided Generative Semantic Enhancement and Disentangled Adaptive Fusion Network for Multimodal Sentiment Analysis

Kejian Dong

Xinjiang Normal University

Zhenzhen He

hezhenzhen@xjnu.edu.cn

Xinjiang Normal University

Tiquan Gu

Xinjiang University

Yaling Wan

Xinjiang Normal University

Xiaochao Fan

Xinjiang Normal University

Bowen Gu

Xinjiang University

Research Article

Keywords: multimodal sentiment analysis, text-guided semantic enhancement, generative semantic enhancement, disentangled representation learning, adaptive fusion

Posted Date: October 28th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11320648/v1>

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

[Read Full License](#)

Additional Declarations: No competing interests reported.

TGSE-DFN: Text-guided Generative Semantic Enhancement and Disentangled Adaptive Fusion Network for Multimodal Sentiment Analysis

Abstract

Multimodal sentiment analysis (MSA) integrates textual, acoustic, and visual information to recognize human affective states. However, two challenges remain. First, the semantic gap between text and nonverbal modalities limits the extraction of high-level affective cues from audio and vision, while text-dominant methods often underuse their inherent emotional information. Second, existing fusion strategies struggle to jointly capture global sentiment tendencies and fine-grained temporal dynamics, such as prosodic changes, pauses, and facial expressions. To address these issues, we propose a Text-Guided Generative Semantic Enhancement and Disentangled Adaptive Fusion Network (TGSE-DFN). TGSE-DFN uses textual semantics as a cross-modal anchor to guide acoustic and visual features into a continuous latent semantic space, enhancing their representation of sentiment polarity, intensity, and contextual meaning. It then applies common-specific disentanglement to separate shared affective information from modality-private cues, reducing redundancy and conflict across modalities. Finally, a global-local pairwise communication adaptive fusion module captures overall sentiment trends through global bottleneck attention, models local temporal dynamics, and dynamically adjusts modality contributions according to sample-specific cue reliability. Experiments on CMU-MOSI and CMU-MOSEI demonstrate that TGSE-DFN achieves competitive performance across multiple evaluation tasks. Ablation studies further confirm the effectiveness of its main components.

Keywords: multimodal sentiment analysis, text-guided semantic enhancement, generative semantic enhancement, disentangled representation learning, adaptive fusion

1 Introduction

With the development of short-video platforms, social media, and human-computer interaction technologies, human affective expression is increasingly conveyed through multiple modalities, including text, audio, and vision. Multimodal sentiment analysis (MSA) integrates linguistic content, vocal prosody, facial expressions, and body movements to predict sentiment polarity and intensity [Zadeh et al. \(2016, 2018\)](#). Compared

with unimodal sentiment analysis, MSA exploits complementary cues across modalities to provide a more comprehensive understanding of the complexity of affective expression. It therefore has important applications in human-computer interaction, intelligent healthcare, public-opinion analysis, and virtual digital humans. An example of multimodal sentiment analysis is shown in Fig. 1.

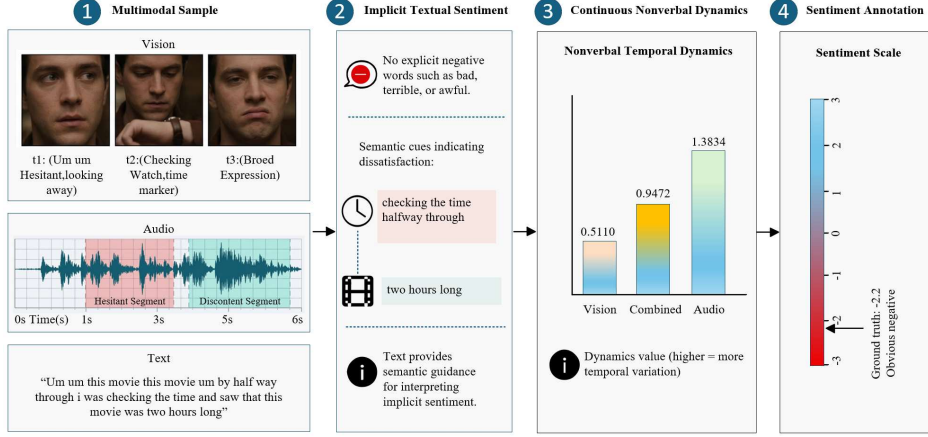


Fig. 1 An example of multimodal sentiment analysis.

In recent years, MSA research has mainly progressed along two directions: cross-modal information fusion and modality representation learning. Representative fusion methods include the Tensor Fusion Network (TFN) proposed by Zadeh et al. Zadeh et al. (2017), which captures high-order interactions among text, audio, and vision through tensor outer products, and the Multimodal Transformer (MulT) proposed by Tsai et al. Tsai et al. (2019), which models dynamic interactions among unaligned modality sequences using cross-modal attention. In addition, low-rank fusion and memory-based fusion methods have also been explored to improve multimodal interaction modeling Liu et al. (2018); Zadeh et al. (2018). To alleviate modality heterogeneity, Hazarika et al. Hazarika et al. (2020) proposed Modality-Invariant and -Specific Representations (MISA), which decomposes multimodal features into shared and modality-specific components. More recently, Ou and Li Ou and Li (2026) introduced the Text-guided Progressive Fusion Network (TPFN), demonstrating that textual semantics can provide effective guidance for cross-modal interaction and fusion. Despite these advances, most existing methods still integrate, align, or constrain multimodal information primarily at the feature level, with limited exploration of the high-level affective semantics implicit in audio and visual signals. Text usually conveys explicit semantic information, whereas audio and vision mainly encode nonverbal affective cues, such as prosody, pauses, facial expressions, and body movements. Without effective semantic guidance, nonverbal modalities may struggle to form representations associated with sentiment polarity, intensity, and contextual meaning, thereby limiting the complementarity of multimodal information.

With the development of pretrained language models and prompt learning, linguistic semantic spaces have increasingly been used to enhance multimodal affective representations [Devlin et al. \(2019\)](#); [Raffel et al. \(2020\)](#). For instance, Wang et al. [Wang et al. \(2026\)](#) proposed Prompt-based Distribution Learning (PDL), which models sentiment-label distributions through prompting to alleviate the subjectivity and ambiguity of sentiment annotations. Recent prompt-efficient adaptation and semantic transfer strategies have further explored the use of pretrained knowledge for multimodal representation learning [Hu et al. \(2022\)](#). Text-enhanced and description-based methods have also introduced linguistic information to improve the semantic expressiveness of nonverbal modalities [Wang et al. \(2023\)](#); [Wu et al. \(2025\)](#); [Liu et al. \(2024\)](#). Other text-guided and cross-modal enhancement frameworks have also investigated semantic interaction between modalities from different perspectives [Yang et al. \(2025\)](#); [He et al. \(2025\)](#). These studies indicate that linguistic semantic knowledge can help reduce the semantic discrepancy between verbal and nonverbal modalities. However, discrete textual descriptions are often summary-level abstractions of acoustic and visual sequences and may discard continuous temporal details, including prosodic variations, pauses, and facial micro-expressions. In addition, the quality of generated descriptions or prompts may directly affect downstream sentiment prediction. Therefore, introducing textual semantic knowledge to enhance acoustic and visual representations without compressing their fine-grained temporal information remains an important problem. Therefore, we use the following example to provide further illustration.

We further conduct a case analysis using a real sample from the MOSI dataset [Zadeh et al. \(2016\)](#). For a segment of a video, the text transcription is “um um this movie this movie um by half way through i was checking the time and saw that this movie was two hours long.” Within the annotation range of $[-3, 3]$, the ground-truth sentiment label of this sample is -2.2 , indicating a clear negative sentiment tendency. The text does not directly contain explicit negative words such as “bad,” “terrible,” or “awful”; instead, it indirectly conveys dissatisfaction through the semantics of “checking the time halfway through” and “a long movie duration.” Further analysis of the preprocessed nonverbal feature sequences shows that the acoustic dynamics value is 1.3834, the visual dynamics value is 0.5110, and the combined nonverbal dynamics value is 0.9472. Here, the dynamics value is measured as the mean norm of feature differences between adjacent time steps and reflects the degree of temporal variation in nonverbal feature sequences. This example indicates that real-world multimodal sentiment samples may exhibit implicit textual sentiment expressions together with continuous temporal changes in nonverbal modalities. Therefore, relying solely on feature-level fusion or compressing nonverbal modalities into discrete textual descriptions may fail to preserve continuous temporal affective information in audio and vision. It is necessary to introduce textual semantic knowledge for enhancement while retaining the temporal structure of nonverbal modalities.

Moreover, multimodal affective expression exhibits clear multi-granular characteristics. A sample contains not only an overall sentiment trend but may also involve key variations in local temporal segments. For example, when a text is implicit or ironic, prosodic changes, pauses, and facial expressions may become decisive cues

for determining sentiment polarity. Tao et al. [Tao et al. \(2025\)](#) proposed the Multi-Focus-Driven Multi-Branch Network (MFMB), which emphasizes the importance of collaborative modeling of global and local information through multi-branch and multi-focus structures. More recently, Zhao et al. [Zhao et al. \(2026\)](#) proposed CCAF, which strengthens the interaction between affective cues at different granularities through coarse-to-fine cross-modal alignment and fusion. Zhang et al. [Zhang et al. \(2026\)](#) proposed SSEDF, which combines shared-private semantic enhancement with dynamic fusion and provides a new perspective on multi-granular affective information modeling. Nevertheless, existing multi-granular fusion methods typically focus more on the fusion architecture itself and less often consider nonverbal semantic enhancement, common-specific representation disentanglement, and sample-level differences in modality contributions simultaneously. Thus, achieving semantic enhancement, representation disentanglement, and multi-granular adaptive fusion within a unified framework remains a key problem for further improving MSA performance.

Based on the above analysis, this paper proposes the Text-Guided Generative Semantic Enhancement and Disentangled Adaptive Fusion Network, TGSE-DFN. Using textual semantics as a cross-modal semantic anchor, the method first guides audio and vision to map into a continuous latent textual semantic space, allowing nonverbal modalities to acquire high-level semantic cues related to sentiment polarity, sentiment intensity, and contextual semantics and thereby enhancing their affective expressiveness. It then feeds the generated semantics back into textual, acoustic, and visual representations to achieve tri-modal collaborative enhancement. Next, common-specific disentanglement is employed to reduce redundancy and conflict across modalities. Finally, a global-local pairwise communication adaptive fusion strategy jointly models overall sentiment trends and local dynamic variations and dynamically adjusts modality contribution weights, thereby improving the stability and discriminability of sentiment prediction.

The main contributions of this paper are summarized as follows:

- 1) We propose a text-guided generative semantic enhancement mechanism that uses textual semantics as a cross-modal semantic anchor to guide audio and vision into a continuous latent textual semantic space. The generated semantics are injected as priors into tri-modal representations, thereby enhancing the high-level affective semantic expression of nonverbal modalities.
- 2) We construct a common-specific disentanglement and global-local adaptive fusion framework. While reducing inter-modality redundancy and conflict, it jointly models cross-modal shared information, modality-private information, global sentiment trends, and dynamic variations along the local temporal dimension, and adaptively adjusts modality contributions according to sample differences.
- 3) We validate the effectiveness of the model on the CMU-MOSI and CMU-MOSEI datasets and analyze the contribution of each key design through ablation studies.

The remainder of this paper is organized as follows. Section 2 reviews research on MSA and text-guided semantic enhancement. Section 3 details the proposed model architecture and training objectives. Section 4 describes the experimental settings and validates the model through performance comparisons, ablation studies, parameter analyses, and visualization analyses. Section 5 concludes the paper and outlines future research directions.

2 Related Work

2.1 Multimodal Sentiment Analysis and Cross-modal Interaction Modeling

Multimodal sentiment analysis (MSA) integrates multi-source information such as text, audio, and vision to predict users’ sentiment polarity and intensity [Zadeh et al. \(2016, 2018\)](#). Compared with unimodal sentiment analysis, MSA exploits complementary information across modalities to capture semantic, prosodic, and behavioral cues in affective expression more comprehensively. However, substantial differences exist across modalities in terms of their forms of expression, temporal scales, and feature distributions. Effectively modeling cross-modal interactions has therefore remained a central problem in this task.

Early MSA approaches primarily modeled multimodal relationships through static feature fusion. For example, Zadeh et al. [Zadeh et al. \(2017\)](#) proposed TFN, which explicitly models high-order interactions among unimodal, bimodal, and trimodal representations through tensor outer products. Liu et al. [Liu et al. \(2018\)](#) further proposed Low-rank Multimodal Fusion (LMF), which preserves multimodal fusion capability while reducing the computational complexity introduced by tensor fusion. Although these approaches can capture global cross-modal correlations, they generally overlook dynamic temporal variations in multimodal sequences.

Subsequently, researchers began to focus on dynamic interaction modeling. Zadeh et al. [Zadeh et al. \(2018\)](#) proposed the Memory Fusion Network (MFN), which captures cross-modal interaction processes through a memory mechanism. Tsai et al. [Tsai et al. \(2019\)](#) proposed MulT, which uses cross-modal attention to process unaligned textual, acoustic, and visual sequences and effectively models long-range dependencies across modalities. Rahman et al. [Rahman et al. \(2020\)](#) proposed MAG-BERT, which injects audio and visual information as nonverbal shifts into BERT representations, thereby enhancing the sensitivity of textual representations to multimodal affective cues. Recent multimodal representation methods have further investigated cross-modal interaction and semantic alignment mechanisms [Yang et al. \(2020\)](#); [Wang et al. \(2023\)](#). More recently, Ou and Li [Ou and Li \(2026\)](#) proposed TPFN to enhance cross-modal interaction modeling through text-guided progressive fusion, while Zhao et al. [Zhao et al. \(2026\)](#) proposed CCAF to further improve interactions among modal information at different granularities through coarse-to-fine cross-modal alignment and fusion. These approaches substantially improve cross-modal interaction modeling; however, most still integrate information primarily in feature space and insufficiently explore the high-level affective semantics implicit in audio and vision.

In recent years, studies have also improved multimodal representation quality from the perspectives of self-supervised learning, contrastive learning, and mutual-information modeling. For example, Yu et al. [Yu et al. \(2021\)](#) proposed Self-MM, which generates unimodal supervision signals in a self-supervised manner to strengthen the semantic association between unimodal and multimodal representations. Han et al. [Han et al. \(2021\)](#) proposed MMIM, which improves the consistency and complementarity of multimodal representations through mutual-information maximization. Liu and Wang [Liu and Wang \(2026\)](#) further proposed FINE, which performs factorized

multimodal sentiment analysis through mutual-information estimation and provides a new modeling perspective for multimodal representation learning. These approaches have advanced MSA from the perspective of representation learning but rarely enhance affective semantics in nonverbal modalities explicitly. Therefore, improving the semantic expressiveness of audio and vision before feature fusion remains an important research question.

2.2 Text-guided Semantic Enhancement and Prompt Learning

The textual modality usually contains more direct and complete affective semantics; consequently, many MSA methods treat text as the core modality. For example, MAG-BERT proposed by Rahman et al. [Rahman et al. \(2020\)](#) injects nonverbal modality information into a pretrained language model, enabling textual representations to perceive acoustic and visual cues. Wang et al. [Wang et al. \(2019\)](#) proposed RAVEN, which dynamically adjusts word-level representations using nonverbal modalities so that linguistic representations can incorporate acoustic and visual changes. These approaches demonstrate the importance of the textual modality in sentiment analysis. However, most emphasize injecting audio and vision as auxiliary information into text, whereas enhancement of the semantic expressiveness of nonverbal modalities themselves remains relatively insufficient.

With advances in pretrained language models and prompt learning, researchers have further attempted to use linguistic semantic spaces to enhance multimodal affective representations [Devlin et al. \(2019\)](#); [Raffel et al. \(2020\)](#). Wang et al. [Wang et al. \(2026\)](#) proposed PDL, which models sentiment-label distributions through prompting to alleviate the subjectivity and ambiguity of sentiment annotations. Ou and Li [Ou and Li \(2026\)](#) proposed TPFN, which enhances semantic interactions among text, audio, and vision through text-guided progressive fusion. Zhao et al. [Zhao et al. \(2026\)](#) proposed CCAF, which performs cross-modal alignment and fusion from coarse to fine, allowing textual semantics to guide the modeling of audio and visual information more effectively. In addition, text-enhanced and description-based methods have been explored to introduce linguistic semantic information into multimodal sentiment representation learning [Wang et al. \(2023\)](#); [Wu et al. \(2025\)](#); [Liu et al. \(2024\)](#). These approaches suggest that mapping nonverbal modalities into a textual semantic space can reduce semantic discrepancies across modalities.

However, existing text-guided methods still have limitations. On the one hand, discrete textual descriptions are generally summary-level expressions of audio and visual signals and may discard continuous temporal details, such as prosodic changes, pauses, and facial micro-expressions. On the other hand, the quality of generated descriptions directly affects downstream sentiment prediction; if a description deviates from genuine affective cues, it may introduce additional noise. Unlike these approaches, this work does not directly generate natural-language descriptions. Instead, it uses a pretrained generative encoder to produce continuous latent textual semantic representations for audio and vision. These representations can both inherit affective knowledge from the textual semantic space and preserve the sequential structure of nonverbal modalities, making them more suitable for subsequent cross-modal semantic injection and disentangled fusion.

2.3 Disentangled Representation Learning

Multimodal data contain both cross-modal shared information and modality-specific information. Shared information generally corresponds to affective factors jointly expressed by different modalities, whereas modality-specific information contains complementary cues unique to a modality, such as sentiment words in text, prosodic changes in audio, and facial movements in vision. Audio and visual affective analysis has also benefited from specialized feature representations, including acoustic descriptors and facial expression analysis tools [Degottex et al. \(2014\)](#); [Stöckli et al. \(2018\)](#). Indiscriminately mixing these types of information during fusion can introduce modality redundancy or even modality conflict, thereby affecting final sentiment prediction [Hazarika et al. \(2020\)](#); [Wang et al. \(2025\)](#). Information-theoretic and statistical dependency analysis methods also provide useful perspectives for separating shared and modality-specific representations [Gretton et al. \(2005\)](#); [Kornblith et al. \(2019\)](#).

To address this problem, many studies have focused on shared-private representation learning. [Hazarika et al. \(2020\)](#) proposed MISA, which partitions multimodal representations into shared and private spaces using modality-invariant and modality-specific constraints to alleviate modality heterogeneity. [Wang et al. \(2025\)](#) proposed DLF, which guides complementary information from nonverbal modalities into linguistic representations from a language-focused perspective and enhances language-dominant affective expression through a disentanglement mechanism. More recently, [Jian et al. \(2026\)](#) proposed MDF, which performs disentanglement and fusion from a modality-aware perspective to improve the modeling of shared and specific information across modalities. [Zhang et al. \(2026\)](#) proposed SSEDF, which combines shared-private semantic enhancement with dynamic fusion to reduce modality redundancy and improve the discriminability of fused representations. [Liu and Wang \(2026\)](#) proposed FINE, which performs factorized multimodal affective modeling through mutual-information estimation to further enhance the complementarity of multimodal representations. These studies demonstrate that appropriately distinguishing shared information from modality-private information can improve MSA performance. Feature optimization strategies, including focal loss based learning, have also been explored to improve representation discrimination in learning systems [Lin et al. \(2017\)](#).

However, existing disentanglement methods still face several limitations. First, most methods perform shared-private decomposition directly in the original feature space. Because a substantial semantic gap exists between acoustic or visual features and textual representations, the learned shared space may lack sufficient high-level affective semantics, while the modality-specific space may fail to adequately preserve fine-grained nonverbal affective cues. Second, shared and private information are not always strictly separable in real-world affective expression. The same prosodic or facial cue may simultaneously convey cross-modal sentiment consistency and modality-specific complementary information. Therefore, excessively strong orthogonality, adversarial, or independence constraints may force potentially useful affective information into mutually exclusive spaces, resulting in information loss. Third, existing methods commonly adopt relatively static disentanglement mechanisms and apply

similar decomposition criteria to different samples. However, the reliability and contribution of each modality vary across samples, and noisy or ambiguous modality features may still be retained in the private space or leak into the shared space. These limitations may weaken the semantic completeness and discriminability of the disentangled representations.

In addition, affective expression generally exhibits clear multi-granular characteristics. Global sentiment trends reflect the overall affective state of an utterance segment, whereas local prosodic pauses, facial changes, or semantic shifts may have a decisive influence on the final sentiment judgment. Tao et al. [Tao et al. \(2025\)](#) proposed the Multi-Focus-Driven Multi-Branch Network (MFMB), which jointly models global and local affective information through multi-branch and multi-focus structures and emphasizes the synergistic role of affective cues at different levels. Zhao et al. [Zhao et al. \(2026\)](#) proposed CCAF, which enhances modal interactions at local and global levels through coarse-to-fine cross-modal alignment and fusion. Some hierarchical Transformer and adaptive fusion approaches have also attempted to characterize local and global dependencies through multi-layer attention or hierarchical interaction structures [Vaswani et al. \(2017\)](#); [Gao et al. \(2026\)](#). Recent studies have further explored fine-grained interaction, semantic constraints, and adaptive distribution modeling for multimodal sentiment analysis [Lin et al. \(2025\)](#); [Xu et al. \(2025\)](#); [Yang et al. \(2024\)](#); [Zhi et al. \(2024\)](#); [Li et al. \(2025\)](#); [Li et al. \(2025\)](#); [Zhu et al. \(2024\)](#). Nevertheless, existing multi-granular fusion approaches typically focus more on the fusion architecture itself and less often consider nonverbal semantic enhancement, common-specific representation disentanglement, and sample-level modality contribution differences simultaneously.

In summary, existing MSA methods have made important progress in cross-modal interaction, text-semantic enhancement, disentangled representation learning, and multi-granular fusion. Nevertheless, current disentanglement methods remain limited by insufficient semantic guidance, unclear boundaries between shared and private information, and limited adaptability to sample-specific modality reliability. A unified framework that simultaneously achieves continuous semantic enhancement for nonverbal modalities, tri-modal collaborative enhancement, common-specific disentanglement, and multi-granular adaptive fusion is still lacking. The proposed TGSE-DFN is designed to address these limitations.

3 Method

This section introduces the proposed Text-Guided Generative Semantic Expert and Disentangled Adaptive Fusion Network, TGSE-DFN. The overall architecture is shown in Fig. 2.

3.1 Problem Definition and Overall Framework

This study uses preprocessed multimodal features as model inputs. Specifically, the text modality is encoded with BERT, acoustic features are extracted from the audio modality using COVAREP, and facial expression and behavioral features are extracted from the visual modality via Facet.

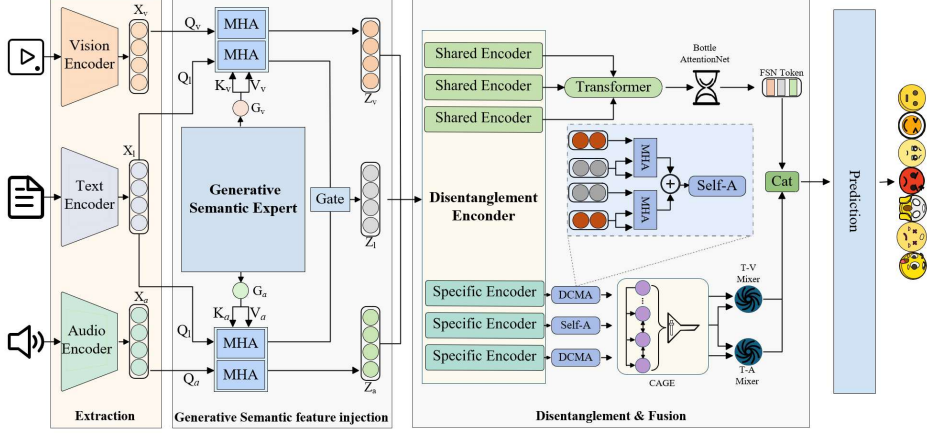


Fig. 2 Overall architecture of TGSE-DFN.

Given a multimodal sample $X = \{X_l, X_a, X_v\}$, where $X_l \in \mathbb{R}^{B \times L_l \times d_l}$ denotes the textual input, $X_a \in \mathbb{R}^{B \times L_a \times d_a}$ denotes acoustic sequence, and $X_v \in \mathbb{R}^{B \times L_v \times d_v}$ denotes visual sequence, respectively, the objective of MSA is to predict continuous sentiment intensity:

$$\hat{y} = F(X_l, X_a, X_v), \quad \hat{y} \in [-3, 3] \quad (1)$$

The overall TGSE-DFN pipeline comprises five stages:

1. Text-guided generative semantic expert;
2. Generated-semantics-driven tri-modal collaborative enhancement;
3. Modality-aware common-specific disentanglement;
4. Text-centered bidirectional specific cross-modal interaction;
5. Global-local pairwise communication adaptive fusion and sentiment prediction.

3.2 Text-guided Generative Semantic Expert

Audio and visual modalities usually convey affect through continuous signals, and their high-level semantics differ substantially from textual semantics. To enhance the affective semantic expression of nonverbal modalities, we design a T5-encoder-based text-guided generative semantic expert. This module does not generate discrete natural-language descriptions; instead, it produces continuous latent semantic representations for audio and vision.

3.2.1 Text Encoding and Non-verbal Temporal Projection

The textual input is mapped into the T5 hidden space:

$$U_l = \text{LN}(\text{GELU}(W_l X_l + b_l)) \quad (2)$$

Audio and visual modalities exhibit clear temporal structures, and different temporal scales may correspond to different affective variations. We use a multi-scale

convolutional projection module to encode nonverbal modalities. For modality $m \in \{a, v\}$, the projection is defined as follows:

$$U_m = \text{LN}\left(\text{Dropout}\left(\text{Conv1D}_1\left(\text{GELU}\left(\text{Cat}_{k \in \mathcal{K}} \text{Conv1D}_k(X_m)\right)\right)\right) + W_s X_m\right) \quad (3)$$

Here, $\mathcal{K} = \{1, 3, 5, 7\}$; W_s denotes residual mapping, and $\text{LN}(\cdot)$ denotes layer normalization. This module extracts local affective variations in audio and vision while preserving sequence structure.

3.2.2 Text-guided Cross-attention Gating

Affective cues in audio and vision are often implicit, and directly feeding them into the T5 encoder may not emphasize components relevant to textual affective semantics. Therefore, to select parts of audio and vision associated with textual affective semantics, we design a text-guided cross-attention gating module. It aligns and filters nonverbal modalities with textual semantics before they enter the T5 encoder.

Taking audio as an example, audio is first used as the Query and text as the Key and Value. Cross-modal attention calculates the relevance between acoustic temporal segments and textual semantic units, thereby capturing correspondences between acoustic affective cues and textual affective semantics. Specifically, this process enables the model to attend to sentiment words, degree expressions, negation, and contextual semantic information associated with acoustic features such as prosodic variation, pauses, and stress. For modality $m \in \{a, v\}$, the text-guided nonverbal representation is formulated as:

$$\hat{U}_m = U_m \odot \left[\rho + (1 - \rho) \sigma \left(\text{MLP}([U_m; \text{LN}(\text{MHA}(U_m, U_l, U_l))]) \right) \right], \quad m \in \{a, v\} \quad (4)$$

The notation $\text{MHA}(\cdot)$ denotes multi-head attention. In the accompanying expression, $[\cdot; \cdot]$ and $\sigma(\cdot)$ represent feature-dimension concatenation and the Sigmoid function, respectively. ρ is a learnable residual weight, and $\odot$ denotes element-wise multiplication.

3.2.3 Prompt-driven Non-verbal Semantic Enhancement

To exploit semantic knowledge in pretrained language models for enhancing nonverbal modalities, we construct a prompt-driven nonverbal semantic enhancement module based on the T5 encoder. Although the original T5 architecture comprises an encoder and a decoder and is commonly used for text-to-text generation, the goal here is not to produce readable natural-language descriptions but to construct continuous latent semantic representations relevant to the sentiment task for audio and vision. We

therefore use only the T5 encoder as a semantic knowledge extractor and guide it to attend to affect-relevant semantic information through task hard prompts and learnable soft prompts. Compared with autoregressively generating discrete text through the T5 decoder, the continuous hidden states output by the encoder can participate more directly in subsequent feature alignment and fusion, while avoiding the compression of nonverbal temporal information by discrete text descriptions and the effect of generation errors on sentiment prediction. The module is illustrated in Fig. 3.

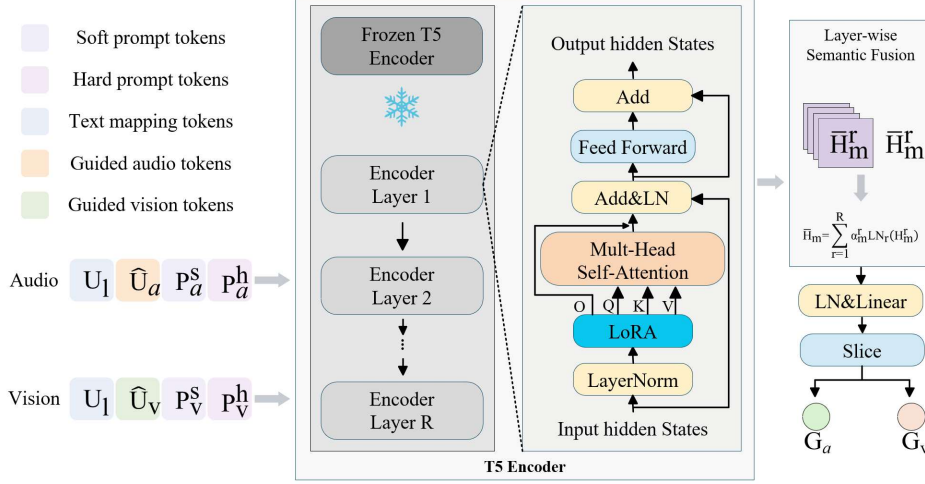


Fig. 3 Architecture of the prompt-driven nonverbal semantic enhancement module.

We construct a fixed task hard prompt for each of the audio and visual modalities. The task hard prompt consists of a task description, modality-analysis instructions, a sentiment-scale specification, and modality-specific semantic-slot prompts. Specifically, the audio semantic slots include Acoustic_Cues, Text_Context, and Sentiment_Level, corresponding to acoustic affective cues, contextual textual tone, and sentiment level, respectively. The visual semantic slots include Visual_Cues, Text_Context, and Sentiment_Level, corresponding to visual affective cues, contextual textual tone, and sentiment level, respectively. The task hard prompt is transformed into the hard-prompt embedding P_m^h through T5 tokenization and the shared word-embedding layer. In addition, a set of learnable continuous prompt vectors P_m^s is introduced as soft prompts for each nonverbal modality. The T5 input sequence is formulated as:

$$Z_m = [U_l; \hat{U}_m; P_m^s; P_m^h], \quad m \in \{a, v\} \quad (5)$$

Here, P_m^h denotes the hard-prompt embedding, and P_m^s denotes the learnable soft-prompt vectors. The hard prompt serves as a task-semantic prior, while the soft prompts provide modality-specific task-adaptation information for the frozen T5 encoder.

To reduce training cost, we freeze the main T5 parameters and introduce low-rank adaptation (LoRA) only into the Query, Key, Value, and Output projection matrices

of selected encoder attention layers. Specifically, for an input hidden state X , the original linear mapping and the LoRA-adapted mapping are formulated as follows:

$$\begin{aligned} Y_0 &= XW, \\ Y &= XW + \frac{\alpha}{d_L} XAB \end{aligned} \quad (6)$$

Here, W denotes frozen pretrained parameters, A and B are trainable low-rank matrices, d_L denotes the rank of the low-rank decomposition, and α is a scaling factor. This design adapts the Query, Key, Value, and Output projection matrices to the sentiment task with a small number of trainable parameters while keeping the main pretrained-model parameters fixed.

For the multi-head self-attention layer in the T5 encoder, given the input hidden state H_m^{l-1} at layer l , where $m \in \{a, v\}$ denotes the acoustic or visual semantic-enhancement branch, Query, Key, and Value are first generated through LoRA-adapted linear projections:

$$\begin{aligned} Q_m^l &= H_m^{l-1} W_Q^l + \frac{\alpha}{d_L} H_m^{l-1} A_Q^l B_Q^l, \\ K_m^l &= H_m^{l-1} W_K^l + \frac{\alpha}{d_L} H_m^{l-1} A_K^l B_K^l, \\ V_m^l &= H_m^{l-1} W_V^l + \frac{\alpha}{d_L} H_m^{l-1} A_V^l B_V^l \end{aligned} \quad (7)$$

The resulting Q_m^l , K_m^l , and V_m^l are then fed into multi-head self-attention to calculate semantic dependencies among different tokens within the sequence:

$$Att_m^l = \text{softmax} \left(\frac{Q_m^l (K_m^l)^T}{\sqrt{d_k}} \right) V_m^l \quad (8)$$

After concatenation, the multi-head attention result is passed through the LoRA-adapted Output projection matrix to obtain the output of the attention layer:

$$O_m^l = Att_m^l W_O^l + \frac{\alpha}{d_L} Att_m^l A_O^l B_O^l \quad (9)$$

Finally, residual connection, layer normalization, and a feed-forward network yield the output hidden state of the l -th encoder layer:

$$\begin{aligned} H_m^l &= \text{FFN} \left(\text{LN} (H_m^{l-1} + O_m^l) \right) \\ &\quad + \text{LN} (H_m^{l-1} + O_m^l) \end{aligned} \quad (10)$$

After the above computations are performed sequentially for R encoder layers, the T5 encoder outputs multi-layer hidden states. To make full use of semantic information at different layers, we perform layer-wise semantic fusion and obtain the generated

semantic features as follows:

$$\begin{aligned} \{H_m^r\}_{r=1}^R &= \text{Enc}_{\text{T5}}(Z_m), \\ G_m &= \text{Slice}_m\left(\text{Linear}\left(\text{LN}\left(\sum_{r=1}^R \alpha_m^r \text{LN}_r(H_m^r)\right)\right)\right) \end{aligned} \quad (11)$$

Here, α_m^r denotes the learnable normalized weight of the r -th T5 encoder layer. $\text{Slice}_m(\cdot)$ denotes the hidden representations at the corresponding acoustic or visual positions extracted from the T5 output sequence. The final outputs are the generated acoustic semantics G_a and the generated visual semantics G_v .

3.3 Generated Semantic-driven Tri-modal Collaborative Enhancement

The generated semantic features G_a and G_v contain acoustic and visual affective information extracted from the T5 semantic space. We further inject them into textual, acoustic, and visual representations through cross-modal attention to achieve tri-modal collaborative enhancement.

For an original modality representation X_m and generated semantics G_n , cross-modal semantic injection is defined as:

$$\begin{aligned} \text{CA}(X_m, G_n) &= \text{LN}\left(W_q X_m \right. \\ &\quad \left. + \text{MHA}(W_q X_m, W_k G_n, W_v G_n)\right) \end{aligned} \quad (12)$$

The textual modality simultaneously receives generated acoustic semantics and generated visual semantics. To adaptively balance the contributions of the two types of nonverbal semantics, we design a gating weight and obtain the enhanced textual representation:

$$\begin{aligned} \Gamma &= \sigma\left(W_\gamma[\text{CA}(X_l, G_a); \text{CA}(X_l, G_v)]\right), \\ E_l &= \text{LN}\left(\Gamma \odot \text{CA}(X_l, G_a) \right. \\ &\quad \left. + (1 - \Gamma) \odot \text{CA}(X_l, G_v)\right) \end{aligned} \quad (13)$$

Audio and vision are enhanced through their respective generated semantics, and each modality then passes through a residual enhancement layer:

$$\begin{aligned} E_a &= \text{CA}(X_a, G_a), \\ E_v &= \text{CA}(X_v, G_v), \\ Z_m &= \text{LN}(E_m + \text{MLP}(E_m)), \quad m \in \{l, a, v\} \end{aligned} \quad (14)$$

Unlike language-centered approaches that only inject nonverbal information into text, this method feeds generated semantics back to textual, acoustic, and visual representations simultaneously. This design avoids excessively strengthening the textual modality and enables tri-modal collaborative enhancement.

3.4 Common-specific Disentanglement after Semantic Enhancement

3.4.1 Common-Specific Disentanglement

The enhanced tri-modal representations Z_l , Z_a , and Z_v contain both shared affective factors and modality-specific factors. To preserve shared affective cues while reducing modality redundancy and conflicting information, we perform common-specific disentanglement.

An independent Transformer encoder is used for each modality to extract its specific representation. Common representations are extracted through a shared Transformer encoder. To prevent modality differences from being erased by complete sharing, we introduce modality adapters that apply modality-conditional corrections in the shared common space. This design allows different modalities to align with a unified affective space while preserving their own distributional characteristics:

$$\begin{aligned} S_m &= E_s^m(Z_m), \\ C_m &= E_c(Z_m) + A_m(E_c(Z_m)), \quad m \in \{l, a, v\} \end{aligned} \quad (15)$$

Here, S_m denotes the modality-specific representation, C_m denotes the common representation, E_s^m denotes the modality-specific encoder, E_c denotes the shared encoder, and A_m denotes the modality adapter for modality m . The adapter is implemented by a lightweight nonlinear mapping with GELU activation and layer normalization.

To prevent information loss during disentanglement, we use an enhancement decoder to reconstruct the original enhanced features from common and specific representations. In addition, to calculate the consistency of common representations across modalities, the common representation C_m is flattened and projected into a unified space:

$$\begin{aligned} \hat{Z}_m &= D_m([S_m; C_m]), \\ R_m &= W_m \text{vec}(C_m) + b_m, \quad m \in \{l, a, v\} \end{aligned} \quad (16)$$

Here, $\text{vec}(\cdot)$ denotes vectorization, i.e., flattening a sequence representation into a vector. The reconstruction and common projections above are used to construct disentanglement constraints in the training objective.

3.4.2 Disentanglement Constraint Loss

The disentanglement constraints are used to ensure effective separation between common and specific representations. The loss jointly considers feature reconstruction, common-representation consistency, and independence between common and specific

representations:

$$\begin{aligned}\mathcal{L}_{\text{dec}} = & \frac{1}{|\mathcal{M}|} \sum_{m \in \mathcal{M}} \left\| \hat{Z}_m - Z_m \right\|_2^2 + \frac{1}{|\mathcal{P}|} \sum_{(i,j) \in \mathcal{P}} \|R_i - R_j\|_2^2 \\ & + \frac{1}{|\mathcal{M}|} \sum_{m \in \mathcal{M}} \left[\cos^2(\text{vec}(S_m), \text{vec}(C_m)) \right. \\ & \left. + \text{HSIC}(S_m, C_m) \right]\end{aligned}\quad (17)$$

Here, $\mathcal{M} = \{l, a, v\}$ denotes the modality set, and $\mathcal{P} = \{(l, a), (l, v), (a, v)\}$ denotes the set of modality pairs. $\cos(\cdot, \cdot)$ denotes cosine similarity, and $\text{HSIC}(\cdot, \cdot)$ denotes the Hilbert-Schmidt Independence Criterion, which measures the statistical correlation between the two types of representations.

The disentanglement constraint loss $\mathcal{L}_{\text{dec}}$ constrains the relationship between common and modality-specific representations. It guides the common branch to learn cross-modally consistent affective factors while allowing the specific branch to retain affective cues unique to each modality. If only the main task loss is used, the two branches may learn similar or redundant mixed representations, weakening the disentanglement effect. Therefore, this work employs $\mathcal{L}_{\text{dec}}$ as an auxiliary regularizer to reduce information overlap between common and specific representations and improve the stability and discriminability of disentangled representations.

The ablation results show that removing disentanglement-related constraints degrades model performance. Although $\mathcal{L}_{\text{dec}}$ is not the sole determinant of performance improvement, it effectively improves the quality of representation disentanglement and, together with the subsequent fusion module, enhances overall model performance.

3.5 Text-centered Bidirectional Specific Cross-modal Interaction

Modality-specific representations contain affective cues unique to each modality. To capture complementary relationships between text and nonverbal modalities, we design bidirectional specific cross-modal interactions.

For each nonverbal modality $m \in \{a, v\}$, text-to-modality and modality-to-text interactions are modeled jointly as:

$$\begin{aligned}H_{lm} = & \frac{1}{2} \left[\text{Trans}_{l,m}(S_l, S_m, S_m) \right. \\ & \left. + \text{Trans}_{m,l}(S_m, S_l, S_l) \right], \quad m \in \{a, v\}\end{aligned}\quad (18)$$

The interactions are then further enhanced through memory encoders:

$$\begin{aligned}\tilde{S}_l &= \text{Trans}_l^{\text{mem}}(S_l), \\ \tilde{S}_m &= \text{Trans}_m^{\text{mem}}(H_{lm}), \quad m \in \{a, v\}\end{aligned}\quad (19)$$

This module enables bidirectional information exchange between text and audio, as well as between text and vision, thereby enhancing the complementarity of modality-specific affective cues.

3.6 Global-Local Pairwise Communication Adaptive Fusion

Multimodal affective expression generally involves overall sentiment trends, fine-grained variations in local temporal segments, and complementary relationships across modalities. A single fusion strategy is insufficient to integrate all of this information adequately. Accordingly, we design the Global-Local Pairwise Communication Adaptive Fusion (GLPC-AF) module. It uses both common and specific representations to jointly model overall sentiment trends, local temporal variations, and modality-contribution differences across samples. Architecture of the global-local pairwise communication adaptive fusion module is shown in Fig. 4.

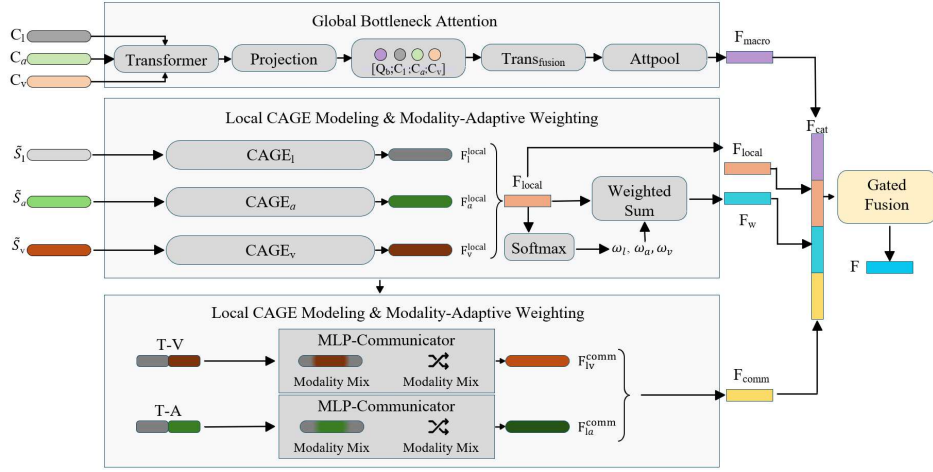


Fig. 4 Architecture of the global-local pairwise communication adaptive fusion module.

3.6.1 Global Bottleneck Attention Fusion

First, the common representations of the three modalities are projected into a unified dimensionality and fed into a bottleneck-attention network. The network assigns an independent Transformer encoder to each modality and introduces learnable bottleneck tokens. Attention pooling is then used to obtain a global macro representation:

$$F_{\text{macro}} = \text{AttPool}\left(\text{Trans}_{\text{fusion}}\left([Q_b; C_l; C_a; C_v]\right)\right) \quad (20)$$

Here, Q_b denotes learnable bottleneck tokens. This branch primarily captures overall sentiment trends from the common information of the three modalities.

3.6.2 Local Temporal Information Modeling

For specific representations, we use CAGE to encode local temporal affective information:

$$F_{\text{local}} = [\text{CAGE}(\tilde{S}_l, M_l), \text{CAGE}(\tilde{S}_a, M_a), \text{CAGE}(\tilde{S}_v, M_v)] \quad (21)$$

Here, M_m denotes a valid-position mask. CAGE consists of a bidirectional gated recurrent unit (BiGRU), convolutional gating, and attention pooling. The BiGRU models temporal dependencies, convolutional gating selects key local variations, and attention pooling aggregates important temporal segments. Therefore, this branch can capture local affective cues such as prosodic variations, pauses, and facial-expression changes.

3.6.3 Pairwise Modality Communication

To further model lightweight pairwise communication between text and nonverbal modalities, we introduce an MLP-Communicator. The MLP-Communicator performs residual mixing along the modality and channel dimensions to capture local complementary relationships between text and nonverbal modalities. The resulting pairwise communication representation is defined as:

$$F_{\text{comm}} = [\text{MLPComm}(F_l^{\text{local}}, F_a^{\text{local}}); \text{MLPComm}(F_l^{\text{local}}, F_v^{\text{local}})] \quad (22)$$

This module captures local complementary relationships between text and audio, as well as between text and vision, at a relatively low computational cost.

3.6.4 Modality-Adaptive Weights

The contributions of modalities vary across samples. We use a modality-adaptive weighting module to dynamically estimate the importance of text, audio, and vision:

$$\omega = \text{softmax}(W_\omega[P_l F_l^{\text{local}}, P_a F_a^{\text{local}}, P_v F_v^{\text{local}}]), \quad (23)$$

$$F_w = \sum_{m \in \{l, a, v\}} \omega_m P_m F_m^{\text{local}}$$

Here, ω denotes the adaptive weights of the three modalities, and P_l , P_a , and P_v denote the modality projection matrices. This mechanism enables the model to select more useful modality information adaptively according to sample content.

3.6.5 Multi-Granular Gated Fusion

The final feature is obtained by concatenating local representations, interaction representations, gated fusion representations, macro representations, and modality-adaptive

representations. The multi-granular gated fusion is formulated as:

$$\begin{aligned} F_{\text{cat}} &= [F_{\text{local}}; F_{\text{comm}}; F_g; F_{\text{macro}}; F_w], \\ F &= F_{\text{cat}} \odot \sigma(\text{MLP}(F_{\text{cat}})) + F_{\text{cat}} \end{aligned} \quad (24)$$

Here, F_g denotes the gated fusion result obtained from concatenated projected shared modality features.

3.7 Sentiment Prediction

After global-local pairwise communication adaptive fusion, the final fused representation F is obtained. It is subsequently fed into a shared prediction network and a task-specific prediction layer to yield the final sentiment prediction:

$$\hat{y} = H_{\text{pred}}(F) \quad (25)$$

3.8 Training Objectives

3.8.1 Task Prediction Loss

Continuous sentiment-intensity prediction is used as the primary optimization objective. Given the ground-truth sentiment label y and the model prediction $\hat{y}$, the L1 loss is used to calculate the task prediction error:

$$\mathcal{L}_{\text{task}} = \frac{1}{B} \sum_{i=1}^B |\hat{y}_i - y_i| \quad (26)$$

Here, B denotes the batch size. This loss directly constrains the discrepancy between the final prediction and the ground-truth sentiment intensity.

3.8.2 Auxiliary Classification Loss

In addition to the main regression task, we introduce an auxiliary binary classification branch to enhance the model’s discriminability at the positive-negative sentiment-polarity boundary. Specifically, continuous sentiment labels are converted into positive and negative classes, and the binary classification branch is supervised using Focal Loss, defined as follows:

$$\mathcal{L}_{\text{focal}} = -\alpha_t(1 - p_t)^\gamma \log(p_t) \quad (27)$$

Here, p_t denotes the predicted probability of the ground-truth class, α_t denotes the class-balancing factor, and γ denotes the focusing parameter. This loss is applied only to the auxiliary binary classification branch to improve discrimination between positive and negative samples and does not replace the main regression loss.

3.8.3 Total Loss

The final training objective is:

$$\mathcal{L} = \mathcal{L}_{\text{task}} + \lambda_{\text{dec}}\mathcal{L}_{\text{dec}} + \lambda_{\text{focal}}\mathcal{L}_{\text{focal}} \quad (28)$$

Here, $\mathcal{L}_{\text{dec}}$ denotes the feature-disentanglement constraint loss, and λ_{dec} and λ_{focal} are the weighting coefficients for the disentanglement loss and Focal Loss, respectively.

4 Experiments

This section systematically evaluates the proposed TGSE-DFN. We first describe the datasets, original feature-extraction methods, model-training settings, and evaluation metrics. We then compare TGSE-DFN with representative existing MSA methods. Finally, the effectiveness of each key module is further verified through ablation experiments, parameter sensitivity analysis, visualization analysis and error analysis.

4.1 Datasets

CMU-MOSI: an English multimodal opinion-level sentiment-intensity dataset containing 2,199 video segments, with sentiment scores ranging from -3 (strongly negative) to 3 (strongly positive).

CMU-MOSEI: a large-scale English multimodal sentiment dataset containing 22,856 utterances, with sentiment scores ranging from -3 (strongly negative) to 3 (strongly positive).

The statistics of the experimental datasets are shown in Table 1.

Table 1 Statistics of the experimental datasets.

Dataset	Language	Sample size	Train	Valid	Test	Modalities	Sentiment
CMU-MOSI	English	2199	1284	229	686	Text, Audio, Vision	$[-3, 3]$
CMU-MOSEI	English	22856	16326	1871	4659	Text, Audio, Vision	$[-3, 3]$

4.2 Experimental Settings and Evaluation Metrics

The model is implemented in PyTorch and trained on aligned multimodal features. The text-guided semantic enhancement module uses a pretrained T5-small encoder as the semantic encoder. Its main parameters are frozen, and only the LoRA parameters, learnable continuous prompt vectors, and downstream task-network parameters are updated. LoRA is injected into the q , k , v , and o projection matrices of the T5 encoder self-attention module, with rank $r = 16$, scaling coefficient $\alpha = 32$, and dropout 0.05. The prompt-vector length for each nonverbal modality is $K = 16$. The model is trained using the AdamW optimizer. For MOSI/MOSEI, the batch size is 32 for both datasets; the BERT learning rates are $1 \times 10^{-5} / 2 \times 10^{-6}$, the T5 adaptation-parameter learning

rates are 3×10^{-5} / 2×10^{-5} , and the learning rates for other modules are 1×10^{-4} / 5×10^{-5} . Text dropout is 0.6 for both datasets, attention dropout is 0.4 for both datasets, and output dropout is 0.6 / 0.1. During training, validation loss is used as the model-selection criterion, and results from the best model are reported on the test sets.

4.3 Baselines Models

To validate the effectiveness of TGSE-DFN, we compare it with several representative MSA methods on the same datasets. The baselines cover conventional fusion methods, cross-modal interaction methods, representation-disentanglement methods, text-centered methods, generative semantic-enhancement methods, and recent multi-granular fusion methods. The baseline models are described below:

MuT(2019) [Tsai et al. \(2019\)](#): a Transformer-based MSA method that uses cross-modal attention to process unaligned textual, acoustic, and visual sequences and explicitly captures long-range inter-modality dependencies.

MAG-BERT(2020) [Rahman et al. \(2020\)](#): a language-centered multimodal pre-training method that injects audio and visual information as nonverbal shifts into BERT representations, enhancing the sensitivity of textual representations to multimodal affective cues.

MMIM(2021) [Han et al. \(2021\)](#): a multimodal representation-learning method based on mutual-information maximization. It improves information association between unimodal and multimodal representations and enhances the discriminability of fused representations.

ALMT(2023) [Zhang et al. \(2023\)](#): a language-guided adaptive hyper-modality Transformer that constructs language-guided hyper-modality representations to strengthen cross-modal interactions among text, audio, and vision.

DEVA(2025) [Wu et al. \(2025\)](#): a vision-audio textualized affective-description method that converts acoustic and visual content into textual affective descriptions to enhance the semantic interpretability and affective expressiveness of nonverbal modalities.

DLF(2025) [Wang et al. \(2025\)](#): a disentangled language-focused MSA method that guides complementary nonverbal information into linguistic representations through common-specific disentanglement and language-focused mechanisms.

MFMB-Net(2025) [Tao et al. \(2025\)](#): a multi-focus-driven multi-branch network that jointly models global and local affective information through a multi-branch architecture and enhances multimodal representation capability through multi-focus fusion.

DIB(2026) [Huang et al. \(2026\)](#): a robust MSA method based on a dual information bottleneck. It compresses unimodal noise and multimodal redundancy through a low-rank Rényi-entropy information bottleneck and learns compact, discriminative multimodal representations through attention-bottleneck fusion.

DGFN(2026) [Han et al. \(2026\)](#): a disentanglement-guided dynamic fusion network that first decomposes modality features into modality-shared and modality-specific representations, then generates sample-level dynamic weights from the disentangled

features, and adaptively fuses multimodal information through a dynamic attention module.

TQ-CHAFN(2026) [Gao et al. \(2026\)](#): a triple-query cross-modal hierarchical adaptive fusion network that treats textual, visual, and acoustic modalities as independent query anchors to construct parallel modality-centric interaction groups, then progressively refines these representations through hierarchical adaptive attention fusion, cross-group consensus modeling, and global gating based on attention-distribution statistics.

The above methods are used as baselines for comparison with TGSE-DFN. Among them, DEVA [Wu et al. \(2025\)](#), DLF [Wang et al. \(2025\)](#), and MFMB-Net [Tao et al. \(2025\)](#) are the most closely related recent representative models: DEVA focuses on textualized semantic descriptions of nonverbal modalities; DLF focuses on language-focused disentangled representation learning; and MFMB-Net focuses on multi-focus, multi-branch fusion. In contrast, TGSE-DFN improves MSA within a unified framework through text-guided generative semantic enhancement, tri-modal semantic injection, common-specific disentanglement, and global-local pairwise communication adaptive fusion.

4.4 Performance Comparison

Table 2 presents experimental results of TGSE-DFN and representative existing methods on the CMU-MOSI and CMU-MOSEI datasets. The best results are shown in bold.

As shown in Table 2, TGSE-DFN outperforms most baseline methods overall on both datasets and achieves competitive performance on the primary evaluation metrics. These results indicate that the proposed text-guided semantic enhancement, shared-specific representation disentanglement, and global-local pairwise communication adaptive fusion mechanisms can effectively improve multimodal sentiment analysis performance. Compared with existing baseline models, the advantages of TGSE-DFN are mainly reflected in the following aspects.

Compared with cross-modal interaction and language-guided methods, including MulT [Tsai et al. \(2019\)](#), MAG-BERT [Rahman et al. \(2020\)](#), and ALMT [Zhang et al. \(2023\)](#), TGSE-DFN consistently achieves better classification performance on both datasets. These methods mainly strengthen multimodal interaction by applying cross-modal attention or injecting nonverbal information into language representations. In contrast, TGSE-DFN explicitly enhances acoustic and visual representations before multimodal fusion. Specifically, textual semantics are used as a cross-modal semantic anchor to guide nonverbal features into a continuous latent semantic space. This design allows acoustic and visual modalities to capture affective cues associated with sentiment polarity, intensity, and contextual meaning, rather than participating in fusion only as low-level auxiliary signals. The performance gains over these interaction-based methods suggest that improving the semantic expressiveness of nonverbal modalities is beneficial for reducing cross-modal heterogeneity.

TGSE-DFN also performs favorably against representation-learning and disentanglement-based methods, such as DLF [Wang et al. \(2025\)](#), MMIM [Han et al. \(2021\)](#), DGFN [Han et al. \(2026\)](#), and DIB [Huang et al. \(2026\)](#). These models

Table 2 Experimental results on the CMU-MOSI and CMU-MOSEI datasets.

Model	CMU-MOSI						CMU-MOSEI					
	Acc7($\uparrow$)	Acc5($\uparrow$)	Acc2($\uparrow$)	F1($\uparrow$)	Corr($\uparrow$)	MAE($\downarrow$)	Acc7($\uparrow$)	Acc5($\uparrow$)	Acc2($\uparrow$)	F1($\uparrow$)	Corr($\uparrow$)	MAE($\downarrow$)
MuT(2019)	40.00	42.68	83.00	82.00	0.698	0.871	51.80	54.18	82.25	82.30	0.703	0.580
MAG-BERT(2020)	43.62	–	84.43	84.61	0.781	0.727	52.67	–	84.82	84.71	0.755	0.543
MMIM(2021)	46.90	53.00	85.30	85.40	0.790	0.712	53.20	55.00	85.00	85.10	0.764	0.536
ALMT(2023)	44.75	50.29	85.37	85.46	0.776	0.738	53.32	55.05	85.94	85.93	0.771	0.534
DEVA(2025)	46.32	51.78	86.29	86.30	0.787	0.730	52.26	55.32	86.13	86.21	0.769	0.541
DLF(2025)	47.08	52.33	85.06	85.04	0.781	0.731	53.90	55.70	85.42	85.27	0.764	0.536
MFMB(2025)	45.80	53.70	85.70	86.00	0.798	0.709	54.20	55.90	85.10	85.10	0.758	0.532
DIB(2026)	47.40	–	85.60	85.60	0.800	0.715	53.50	–	86.00	86.00	0.790	0.588
DGFN(2026)	48.40	53.79	85.67	85.68	0.798	0.717	54.26	56.02	85.58	85.44	0.768	0.533
TQ-CHAFN(2026)	47.96	–	85.82	85.87	0.806	0.701	53.58	–	85.66	85.93	0.775	0.529
TGSE-DFN	48.83	55.98	86.74	86.69	0.799	0.690	55.55	57.39	86.32	86.33	0.774	0.534

improve multimodal representations through information constraints, shared-specific decomposition, dynamic fusion, or information bottlenecks. However, their representation constraints are generally imposed directly on the original modality features. In TGSE-DFN, common-specific disentanglement is conducted after acoustic and visual representations have been semantically enhanced. Therefore, the common branch can capture cross-modal affective consistency in a semantically enriched space, while the specific branches preserve complementary cues unique to individual modalities. This ordering helps prevent the common representation from being dominated by shallow statistical correlations and reduces redundant or conflicting information during subsequent fusion. The improvements in Acc7, Acc5, Acc2, and F1 support the effectiveness of combining semantic enhancement with disentangled representation learning.

The comparison with DEVA Wu et al. (2025) further demonstrates the advantage of continuous latent semantic enhancement. Although DEVA introduces linguistic knowledge by converting acoustic and visual information into textual affective descriptions, discrete descriptions may compress continuous nonverbal variations and are sensitive to generation quality. TGSE-DFN instead uses the hidden representations of the T5 encoder Raffel et al. (2020) without generating natural-language descriptions. Consequently, it can exploit pretrained linguistic knowledge while preserving temporal affective cues such as prosodic changes, pauses, vocal intensity variations, and facial movements. This distinction is particularly relevant to fine-grained sentiment prediction, for which continuous nonverbal variations may provide critical evidence beyond explicit textual content.

Finally, compared with multi-granular and adaptive fusion methods such as MFMB-Net Tao et al. (2025) and TQ-CHAFN Gao et al. (2026), TGSE-DFN achieves stronger overall classification results, particularly for Acc7 and Acc5. MFMB-Net and TQ-CHAFN demonstrate the effectiveness of multi-branch interaction and hierarchical adaptive fusion, but their main emphasis is placed on fusion architecture design. TGSE-DFN performs global-local adaptive fusion on semantically enhanced and disentangled representations. The global branch captures the overall sentiment tendency, whereas the local branch models temporally varying cues in acoustic and visual sequences. Meanwhile, sample-level modality weights allow the model to adjust the contributions of text, audio, and vision according to their reliability in individual samples. Therefore, the proposed fusion strategy can jointly exploit global semantics, local dynamics, and sample-specific modality differences.

Overall, the comparative results validate the three main components of TGSE-DFN: text-guided continuous semantic enhancement improves the affective expressiveness of nonverbal modalities; semantic-space common-specific disentanglement reduces modality redundancy and conflict; and global-local adaptive fusion integrates affective evidence across different temporal granularities and modalities. Their collaboration enables TGSE-DFN to achieve competitive and generally superior performance on both benchmark datasets.

4.5 Ablation Study

To validate the effectiveness of the key TGSE-DFN modules, we conduct ablation studies under the same experimental settings. All ablated models use identical training

Table 3 Ablation results.

Model	MOSI			MOSEI		
	Acc7	Acc2	F1	Acc7	Acc2	F1
w/o T5	47.81	85.21	85.17	54.28	86.08	86.09
w/o Dec	47.96	84.30	84.30	54.39	86.30	86.30
w/o DCMA	47.96	83.54	83.54	53.64	85.72	85.69
w/o MAF	48.25	84.60	84.63	54.86	86.21	86.19
w/o DecLoss	48.83	85.67	85.66	54.60	86.13	86.12
TGSE-DFN	48.83	86.74	86.69	55.55	86.32	86.33

strategies, data splits, and evaluation metrics; only the specified module is removed to analyze its effect on overall performance. The specific ablation settings are as follows:

w/o T5: removes the complete T5 generative semantic enhancement module.

w/o Dec: removes the common-specific disentanglement module.

w/o DCMA: removes the Dual Cross-Modal Attention (DCMA) interaction module for disentangled specific features.

w/o MAF: removes the global-local pairwise communication adaptive fusion module.

w/o DecLoss: removes the disentanglement loss module.

The ablation results are shown in Table 3.:

As shown in Table 3., the complete TGSE-DFN outperforms all ablated variants overall on both datasets, indicating that each key module contributes to performance improvement. After the T5 semantic enhancement module is removed, the Acc-2 and F1 scores of w/o T5 on MOSI decrease to 85.21 and 85.17, respectively. This result indicates that semantic knowledge provided by the pretrained language model can effectively enhance the affective expressiveness of audio and visual modalities.

For disentanglement-related ablations, w/o Dec removes the common-specific disentanglement structure, and model performance declines. This result suggests that the disentanglement mechanism helps distinguish cross-modal shared affective factors from modality-private cues, thereby mitigating modality redundancy and conflict. Furthermore, w/o DecLoss removes the constraint loss while retaining the disentanglement structure, and its performance is also lower than that of the complete model. This finding indicates that structural partitioning alone is insufficient to guarantee a clear division of labor between common and specific representations. The disentanglement constraint loss further reduces information overlap across representation branches and improves the stability and discriminability of disentangled representations.

Further observation shows that the degradation of w/o DCMA is relatively pronounced: its Acc-2 and F1 both fall to 83.54 on MOSI. This result indicates that modality-specific representations after disentanglement still require cross-modal interaction to model complementary relationships among text, audio, and vision. After the global-local pairwise communication adaptive fusion module is removed, w/o MAF also performs worse than the complete model on both datasets, demonstrating that multi-granular fusion can more fully integrate global sentiment trends, local temporal variations, and sample-level differences in modality contributions.

Overall, the ablation results indicate that the performance improvement of TGSE-DFN does not arise from a single module. Rather, it results from the combined effect of T5 semantic enhancement, common-specific disentanglement, the disentanglement constraint loss, cross-modal interaction, and global-local pairwise communication adaptive fusion.

4.6 Parameter Sensitivity Analysis

To analyze the sensitivity of TGSE-DFN to key hyperparameters, we conduct parameter-sensitivity experiments on the MOSI and MOSEI datasets. Three key parameters are analyzed: the number of Transformer fusion layers (nlevels), the length of T5 learnable soft prompts (L), and the Focal Loss focusing parameter (γ). All other parameters remain at their default configuration during each analysis. The default configuration is nlevels=2, $L = 16$, and $\gamma = 2.0$.

The results are shown in Fig. 5–7. The blue curve denotes Acc-2, and the red curve denotes F1. Higher Acc-2 indicates higher binary classification accuracy, whereas higher F1 indicates better overall classification performance on positive and negative sentiment classes. The gray dashed line marks the default parameter values ultimately adopted in this study.

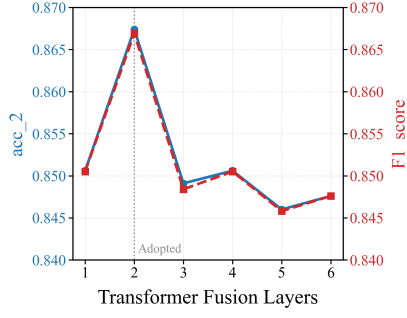
4.6.1 Sensitivity Analysis of the Number of Transformer Fusion Layers

Figs. 5a and 5b show the effect of the number of Transformer fusion layers, nlevels, on model performance. We evaluate nlevels $\in \{1, 2, 3, 4, 5, 6\}$. On MOSI, when the number of fusion layers is 1, Acc-2 and F1 are relatively low, indicating that a single-layer Transformer fusion architecture has limited ability to model complex cross-modal interactions among text, audio, and vision. When the number of fusion layers increases to 2, the model achieves strong performance on both Acc-2 and F1, suggesting that an appropriate increase in fusion depth improves the capture of multimodal contextual relationships and cross-modal complementary information.

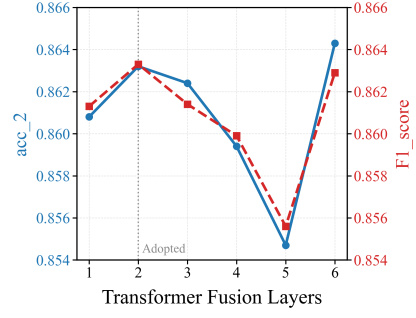
When the number of fusion layers further increases to 3 or more, Acc-2 and F1 on MOSI are generally lower than under the default setting. This finding suggests that an excessively deep fusion structure may introduce more parameters and a more complex optimization process, making overfitting or excessive mixing of modality information more likely on a relatively small dataset. In contrast, because MOSEI is larger and contains richer semantic and modal variations, some deeper settings are slightly better than the default depth on individual metrics, although the overall fluctuation is small, indicating a degree of parameter stability. Considering results on both datasets, we set nlevels=2 to achieve a reasonable balance among cross-modal modeling capability, model complexity, and cross-dataset stability.

4.6.2 Sensitivity Analysis of the Focal Loss Focusing Parameter

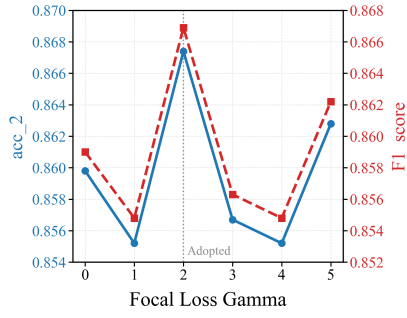
Figs. 5c and 5d show the effect of the Focal Loss focusing parameter γ on model performance. We evaluate $\gamma \in \{0, 1, 2, 3, 4, 5\}$. When $\gamma = 0$, Focal Loss degenerates into



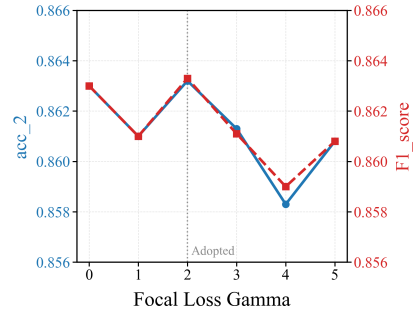
(a) MOSI: Transformer fusion layers.



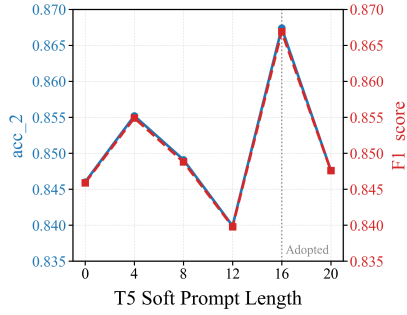
(b) MOSEI: Transformer fusion layers.



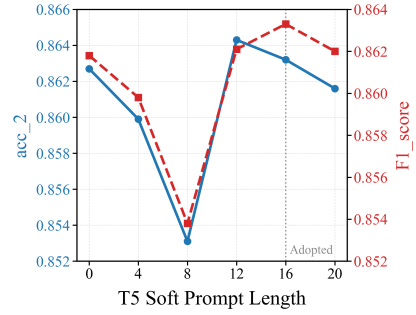
(c) MOSI: Focal Loss focusing parameter γ .



(d) MOSEI: Focal Loss focusing parameter γ .



(e) MOSI: T5 learnable soft-prompt length L .



(f) MOSEI: T5 learnable soft-prompt length L .

Fig. 5 Parameter sensitivity analysis on the CMU-MOSI and CMU-MOSEI datasets: (a)–(b) the number of Transformer fusion layers; (c)–(d) the Focal Loss focusing parameter γ ; and (e)–(f) the T5 learnable soft-prompt length L .

ordinary cross-entropy loss. As γ increases, the loss assigns larger weights to difficult-to-classify samples, thereby improving the model’s attention to difficult samples and samples near class boundaries.

The results show that when $\gamma = 0$, the model already achieves good classification performance. However, because ordinary cross-entropy does not sufficiently distinguish easy from difficult samples, learning for some sentiment-boundary samples remains inadequate. When $\gamma = 1$, performance declines to some extent, indicating that a small focusing parameter may not yet enhance difficult-sample learning effectively and may instead perturb the overall optimization process. When $\gamma = 2$, both Acc-2 and F1 reach their best values, indicating that this setting balances learning from ordinary and difficult samples and enables more stable discriminability in the sentiment classification task.

When γ further increases to 3, 4, or 5, model performance declines or fluctuates to varying degrees. This finding indicates that an excessively large focusing parameter can cause the model to overemphasize a small number of difficult samples, weakening stable learning of the overall sample distribution and potentially amplifying the influence of noisy samples during training. Although performance recovers at some individual values, its overall stability remains inferior to that obtained with $\gamma = 2$. Therefore, considering Acc-2, F1, and training stability, we set the Focal Loss focusing parameter to $\gamma = 2.0$.

4.6.3 Sensitivity Analysis of the T5 Learnable Soft-Prompt Length

Figs. 5e and 5f show the effect of the T5 learnable soft-prompt length L on model performance. We evaluate $L \in \{0, 4, 8, 12, 16, 20\}$. The soft-prompt length determines the number of learnable prompt vectors used during text-guided generative semantic enhancement and directly affects the ability of the pretrained T5 encoder to provide semantic guidance to audio and visual modalities. Here, $L = 0$ indicates that no learnable soft prompt is introduced and semantic enhancement relies only on the fixed textual prompt and original multimodal features.

On MOSI, when $L = 0$, the model can still obtain a degree of semantic enhancement from the fixed textual prompt; however, without learnable prompt vectors, its ability to adaptively map nonverbal modalities into the textual semantic space is limited. When the soft-prompt length increases to $L = 4$ or $L = 8$, performance improves, indicating that a modest number of learnable prompts provides the T5 expert with additional task-relevant semantic guidance. However, performance fluctuates at $L = 12$, suggesting that an unsuitable prompt length may destabilize prompt representations. When the soft-prompt length increases to $L = 16$, Acc-2 and F1 achieve strong performance, indicating that this length provides a reasonable balance between prompt capacity and optimization stability.

When the soft-prompt length further increases to $L = 20$, model performance on MOSI declines or fluctuates. This may be because excessively long soft prompts introduce redundant prompt information, increase optimization difficulty, and may interfere with original textual semantic representations, making it harder for the generative semantic enhancement module to learn effective affect-relevant semantics stably. Unlike MOSI, some longer soft-prompt lengths on MOSEI are slightly better than

the default setting on individual metrics. This may be because MOSEI is larger and contains more diverse affective expressions, allowing longer soft prompts to provide greater learnable semantic capacity. Nevertheless, the overall results remain stable and competitive. Considering performance on both datasets, parameter efficiency, and consistency of experimental settings, the final T5 learnable soft-prompt length is set to $L = 16$.

4.7 Visualization and Qualitative Analysis

To further analyze the internal mechanisms of TGSE-DFN and qualitatively validate the contribution of each module to performance improvement, we conduct visualization analyses on the CMU-MOSI and CMU-MOSEI datasets. Specifically, we examine fused-representation distributions, the effectiveness of common-specific representation disentanglement, T5 semantic enhancement and gating mechanisms, modality-adaptive fusion weights, and fine-grained prediction errors. The corresponding results are shown in Fig. 8–12.

4.7.1 Visualization of the Sentiment Distribution of Fused Representations

Fig. 6 presents t-distributed stochastic neighbor embedding (t-SNE) visualizations of the final fused representations learned by TGSE-DFN on the two datasets. Positive, negative, and neutral samples exhibit a degree of clustering in the low-dimensional space, indicating that the learned fused representations have good sentiment discriminability. At the same time, partial overlap remains between adjacent sentiment classes, especially among mildly negative, neutral, and mildly positive samples, whose boundaries are relatively close. This is consistent with the small differences between adjacent sentiment levels in fine-grained sentiment analysis.

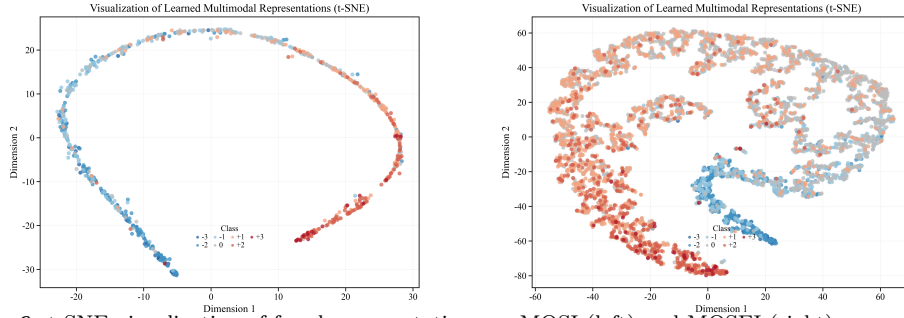


Fig. 6 t-SNE visualization of fused representations on MOSI (left) and MOSEI (right).

Overall, the visualizations on both MOSI and MOSEI indicate that TGSE-DFN can learn multimodal fused representations with sentiment discriminability and a continuous sentiment structure.

4.7.2 Analysis of Common-Specific Representation Disentanglement

Fig. 7 presents centered kernel alignment (CKA) similarities between common and modality-specific representations. Here, S_L , S_A , and S_V denote the modality-specific representations of text, audio, and vision, respectively, whereas C_L , C_A , and C_V denote the corresponding common representations.

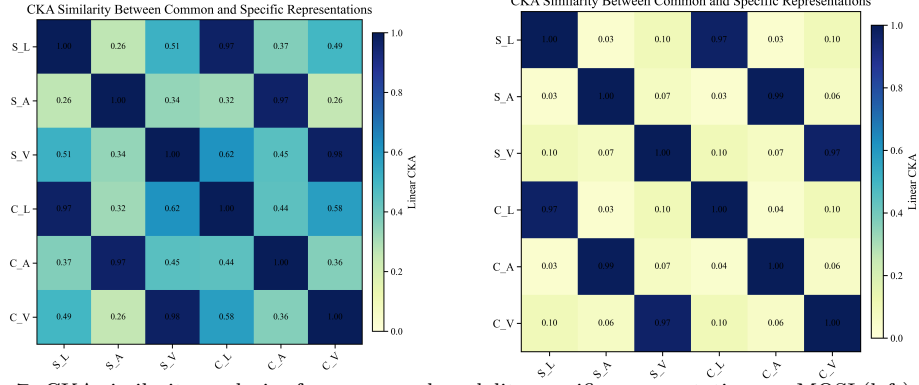


Fig. 7 CKA similarity analysis of common and modality-specific representations on MOSI (left) and MOSEI (right).

The results show that common and specific representations within the same modality retain a certain degree of correlation, indicating that the disentanglement process does not destroy affective information in the original modality. At the same time, similarities among specific representations from different modalities are relatively low, suggesting that the model preserves modality differences among text, audio, and vision. These results indicate that the proposed disentanglement module can effectively distinguish cross-modal shared factors from modality-private cues while retaining affect-relevant information, thereby mitigating modality redundancy and conflict.

4.7.3 Analysis of T5 Semantic Enhancement and the Gating Mechanism

Fig. 8 shows cross-modal attention distributions and gating-coefficient distributions in the T5 semantic enhancement module. Fig. 8(a) and Fig. 8(b) represent the attention responses of textual features to T5-generated acoustic expert features and T5-generated visual expert features, respectively. The attention weights are not uniformly distributed over all generated feature positions; instead, higher responses occur at some positions. This finding suggests that the model can selectively extract useful information from the continuous latent semantic representations generated by the T5 encoder.

Fig. 8(c) further shows the distribution of gating coefficients across sentiment classes. Because the final enhanced representation is computed as Formula 13, a

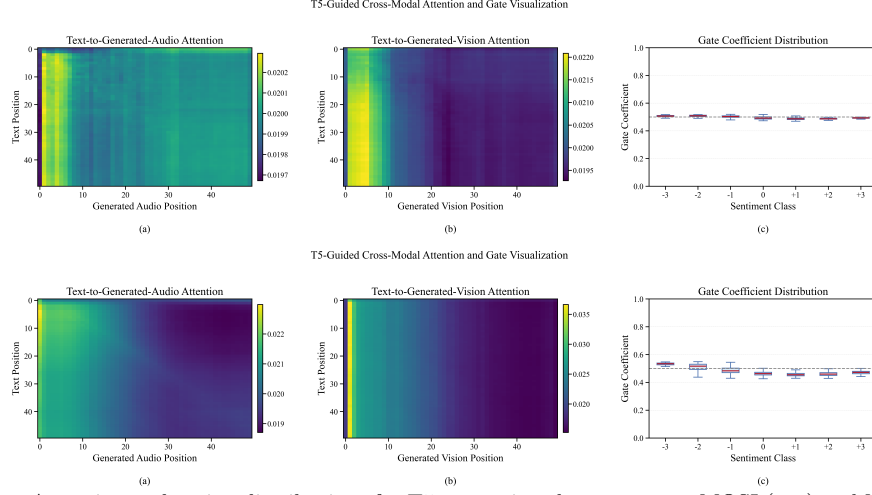


Fig. 8 Attention and gating distributions for T5 semantic enhancement on MOSI (top) and MOSEI (bottom).

gate value greater than 0.5 indicates that the model favors the T5-generated acoustic semantic-enhancement branch, whereas a gate value below 0.5 indicates that it favors the T5-generated visual semantic-enhancement branch. Differences in gating-coefficient distributions across sentiment classes indicate that the model can adaptively balance acoustic and visual semantic enhancement according to sample affective characteristics. This result suggests that the T5 semantic enhancement module provides effective high-level semantic supplementation for nonverbal modalities and dynamically fuses generated expert features from different sources through the gating mechanism.

4.7.4 Analysis of Modality-Adaptive Weights

Fig. 9 shows average fusion weights of the textual, acoustic, and visual modalities across sentiment classes. The modality weights differ across sentiment classes, indicating that TGSE-DFN does not use a fixed fusion ratio but can dynamically adjust modality contributions according to the affective-expression characteristics of each sample.

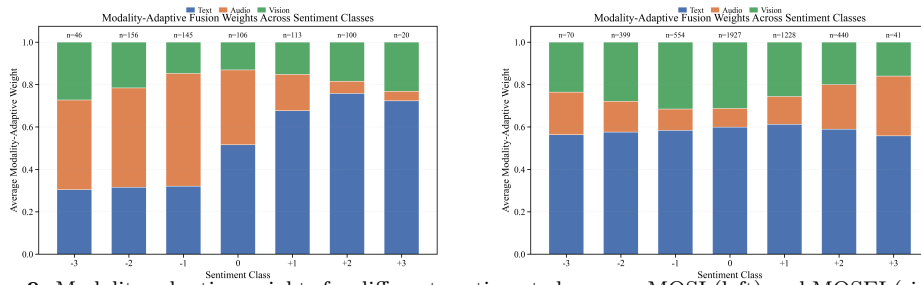


Fig. 9 Modality-adaptive weights for different sentiment classes on MOSI (left) and MOSEI (right).

For some negative sentiment classes, the weights of audio and vision increase relatively, indicating that nonverbal cues such as prosody, pauses, and facial expressions are important for negative-sentiment judgments. For some samples with more explicit affective expression, the textual modality receives higher weights, indicating that the model can fully exploit textual semantic information. This result validates the effectiveness of the global-local pairwise communication adaptive fusion module in modeling sample-level differences in modality contributions.

4.8 Error Analysis

4.8.1 Fine-Grained Prediction Error Analysis

Fig. 10 shows seven-class confusion matrices and class-wise MAE values on the two datasets. The confusion matrices show that model misclassifications are concentrated primarily between adjacent sentiment levels, whereas severe errors that cross polarity are relatively rare. This indicates that TGSE-DFN can judge sentiment polarity reasonably well but still faces challenges in distinguishing fine-grained sentiment intensities.

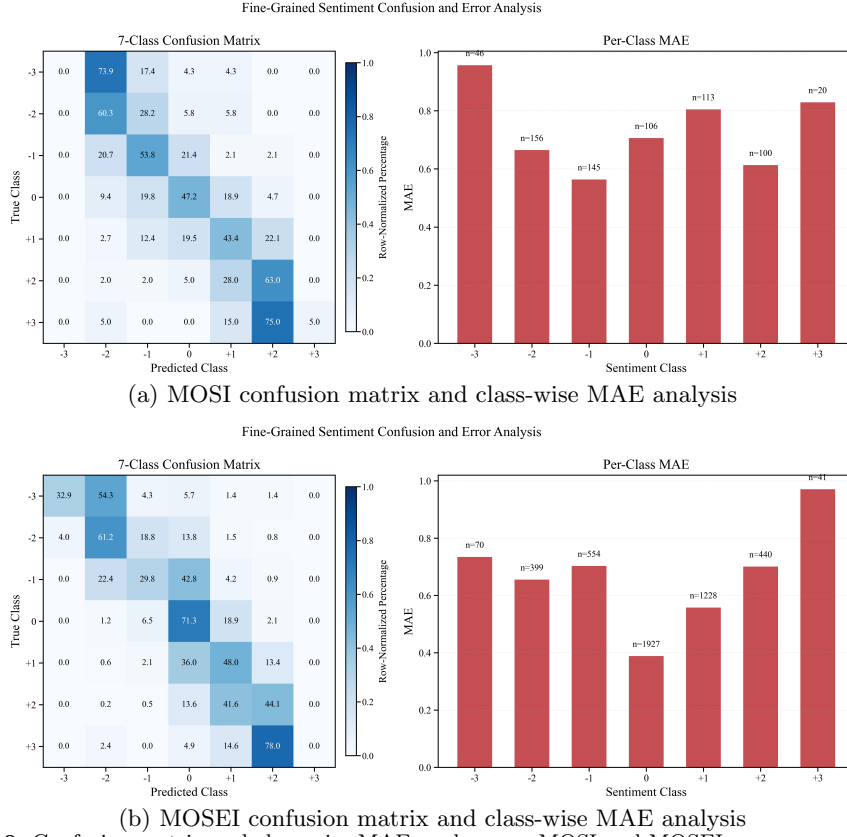


Fig. 10 Confusion-matrix and class-wise MAE analyses on MOSI and MOSEI.

The class-wise MAE values show that prediction errors are relatively high for extreme sentiment classes, possibly because these classes contain fewer samples and have more diverse forms of expression. In contrast, predictions for intermediate sentiment classes are more stable. These results indicate that TGSE-DFN models overall sentiment trends well, but subtle differences between adjacent sentiment intensities remain a major challenge in MSA.

4.8.2 Error Type Distribution and Typical Case Analysis

To further analyze the limitations of the model, we conduct a post hoc error analysis using the best trained model on the MOSI and MOSEI test sets. Specifically, according to the relationship between predictions and ground-truth labels, errors are divided into five typical types: polarity reversal, neutral-boundary error, sentiment-intensity underestimation, sentiment-intensity overestimation, and extreme-sentiment collapse. Polarity reversal refers to predicting the wrong positive/negative sentiment direction. A neutral-boundary error occurs when a ground-truth label is close to neutral but the model predicts a clearly positive or negative value. Sentiment-intensity underestimation and overestimation refer to cases in which the model predicts the correct polarity but underestimates or overstates the sentiment intensity, respectively. Extreme-sentiment collapse refers to predicting a strong-sentiment sample as weakly sentimental or close to neutral.

Table 4 shows the distribution of error types on the two datasets. On MOSI and MOSEI, more than half of samples are categorized as correct predictions or minor errors, accounting for 50.73% and 56.64%, respectively. It should be noted that Correct / Minor in the table is not conventional classification accuracy; it refers to samples for which the predicted sentiment score is close to the ground-truth score. In continuous sentiment regression, even when the model correctly predicts the positive/negative direction, it is categorized as an intensity-underestimation or intensity-overestimation error if the sentiment-intensity estimate deviates substantially. Among the main error types, sentiment-intensity underestimation is the most common on both datasets, accounting for 18.80% on MOSI and 14.83% on MOSEI. This suggests that the model can usually capture basic sentiment polarity but remains insufficient for fine-grained sentiment-intensity estimation. In addition, the polarity-reversal rate on MOSEI is 5.71%, lower than 8.31% on MOSI, suggesting that coarse-grained polarity judgments are more stable on the larger dataset. However, neutral-boundary errors reach 11.89% on MOSEI, higher than 8.16% on MOSI, indicating that the model may still produce overly polarized predictions for neutral or weak-sentiment samples.

Table 5 further presents typical samples for different error types, facilitating more concrete observation of the causes of model errors. For polarity reversal, the MOSI sample “And oh oh my gosh i was blown away” has a ground-truth label of 2.80 but is predicted as -2.35 . Although the text contains the clear strongly positive expression “blown away,” the model produces a negative prediction. This suggests that, for some samples, acoustic prosody, visual expressions, or contextual noise may interfere with the model and cause the multimodal fused representation to deviate from the explicit sentiment direction in text. For a neutral-boundary error, the MOSEI sample “Maybe I can say it was one of the worst days in my life, but it changed to one of the best

Table 4 Statistics of error-type distributions.

Error Type	MOSI Count	MOSI Ratio (%)	MOSI Mean MAE	MOSEI Count	MOSEI Ratio (%)	MOSEI Mean MAE
Polarity Reversal	57	8.31	2.206	266	5.71	1.450
Neutral Boundary	56	8.16	1.036	554	11.89	0.850
Intensity Under-estimation	129	18.80	0.888	691	14.83	0.834
Intensity Over-estimation	67	9.77	0.853	350	7.51	0.821
Extreme Collapse	18	2.62	1.589	92	1.97	1.596
Other Large Error	11	1.60	0.683	67	1.44	0.638
Correct / Minor	348	50.73	0.236	2639	56.64	0.219
Total	686	100.00	–	4659	100.00	–

days in my life later” has a ground-truth label of -0.33 , which is close to neutral, but is predicted as 2.14. The sample contains opposing affective cues, including “worst days” and “best days.” The model favors the positive expression in the latter half and consequently judges the overall sentiment as clearly positive, reflecting insufficient ability to balance mixed sentiments and contrastive semantics.

For sentiment-intensity estimation, the MOSI sample “So if thats the case then you should see this movie if you really like cheesy jokes...” has a ground-truth label of 0.60 but is predicted as 2.18, indicating that the model may overamplify local positive cues such as “really like” and predict weak positive sentiment as strong positive sentiment. Conversely, the model also exhibits underestimation and even collapse for strong-sentiment samples. For example, a MOSEI sample concerning the “Rwandan genocide” has a ground-truth label of -3.00 but is predicted as -1.08 , indicating that the model recognizes the negative polarity but does not fully capture the intense negative affect implicit in the content. Another sample, “If it were a scale of one to ten I’d give it a ten definitely,” has a ground-truth label of 3.00 but is predicted as only 0.02, indicating that the model can compress extreme sentiment toward neutrality for some strongly positive samples. Overall, these examples show that the main limitation of the model is not a complete inability to recognize sentiment direction but insufficiently stable fine-grained modeling of mixed sentiments, neutral boundaries, and strong sentiment intensity. In the future, explicit modal reliability estimation and uncertainty constraints can be further introduced to reduce the interference of noisy modalities and improve the model’s predictive performance on boundary samples and highly emotional samples.

5 Conclusion

To address the insufficient semantic expression of nonverbal modalities, strong modality heterogeneity, and inadequate multi-granular information fusion in MSA, this paper proposes TGSE-DFN. Specifically, the model employs a T5 encoder to construct continuous latent semantic representations for audio and vision, and integrates common-specific representation disentanglement with global-local pairwise communication adaptive fusion to enhance the affective expressiveness of nonverbal modalities and improve the discriminability of multimodal representations. Experimental results on the MOSI and MOSEI datasets demonstrate that TGSE-DFN achieves superior or competitive performance in both sentiment classification and sentiment-intensity prediction. Furthermore, ablation studies validate the effectiveness of semantic enhancement, representation disentanglement, and adaptive fusion, while visualization analyses further show that the proposed model can learn multimodal representations with stronger sentiment discriminability.

This work still has limitations. First, the T5 encoder introduces additional computational overhead. Second, the current prompt design still relies on manual experience, and the model’s generalization to cross-lingual and complex real-world scenarios requires further validation. Future work will explore more lightweight semantic-enhancement architectures and automated prompt-construction methods

Table 5 Analysis of typical error cases.

Dataset	Content	Error Type		Prediction	Label	Main Reason
MOSI	And oh oh my gosh i was blown away	Polarity Reversal		-2.35	2.80	The model predicts the opposite polarity, suggesting that misleading multimodal cues may dominate the fusion representation.
MOSEI	Maybe I can say it was one of the worst days in my life, but it changed to one of the best days in my life later	Neutral Boundary		2.14	-0.33	The utterance contains mixed sentiment expressions, causing the model to over-polarize a near-neutral sample.
MOSEI	Or I see what happened in the Rwandan genocide, with the slamming of babies against the walls of churches...	Intensity estimation	Under-	-1.08	-3.00	The model captures the negative polarity but significantly underestimates the extreme negative intensity.
MOSI	So if thats the case then you should see this movie if you really like cheesy jokes...	Intensity estimation	Over-	2.18	0.60	Weak positive sentiment is over-amplified, leading to an exaggerated positive prediction.
MOSEI	If it were a scale of one to ten I'd give it a ten definitely	Extreme Collapse		0.02	3.00	A strong positive sample is predicted as nearly neutral, indicating severe collapse of sentiment intensity.

and extend this framework to cross-lingual, cross-domain, and larger-scale multimodal affective-understanding tasks.

References

- Devlin, J., Chang, M.-W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of NAACL-HLT, pp. 4171–4186 (2019). <https://doi.org/10.18653/v1/N19-1423>
- Degottex, G., Kane, J., Drugman, T., Raitio, T., Scherer, S.: COVAREP: A collaborative voice analysis repository for speech technologies. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 960–964 (2014). <https://doi.org/10.1109/ICASSP.2014.6853739>
- Gretton, A., Bousquet, O., Smola, A., Schölkopf, B.: Measuring statistical dependence with hilbert-schmidt norms. In: Proceedings of Algorithmic Learning Theory. Lecture Notes in Computer Science, vol. 3734, pp. 63–77 (2005). https://doi.org/10.1007/11564089_7
- Gao, C., Liu, J., Wang, J., Chen, L.: TQ-CHAFN: Triple-query cross-modal hierarchical adaptive fusion network for multimodal sentiment analysis. Knowledge-Based Systems **339**, 115654 (2026) <https://doi.org/10.1016/j.knosys.2026.115654>
- Han, W., Chen, H., Poria, S.: Improving multimodal fusion with hierarchical mutual information maximization for multimodal sentiment analysis. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pp. 9180–9192 (2021). <https://doi.org/10.18653/v1/2021.emnlp-main.723>
- Han, C., Gao, H., Chen, L.: DGFN: Disentanglement-guided dynamic fusion network for multimodal sentiment analysis. Expert Systems with Applications **319**, 132110 (2026) <https://doi.org/10.1016/j.eswa.2026.132110>
- Huang, H., Gong, T., He, K., Wu, J., Cambria, E., Feng, M.: Robust multimodal sentiment analysis via double information bottleneck. Information Fusion **129**, 103964 (2026) <https://doi.org/10.1016/j.inffus.2025.103964>
- He, X., Liang, H., Peng, B., Xie, W., Khan, M.H., Song, S., Yu, Z.: MSAmba: Exploring multimodal sentiment analysis with state space models. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 39, pp. 1309–1317 (2025). <https://doi.org/10.1609/aaai.v39i2.32120>
- Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: LoRA: Low-rank adaptation of large language models. In: Proceedings of the International Conference on Learning Representations (2022)
- 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, pp. 1122–1131 (2020). <https://doi.org/10.1145/3394171.3413678>
- Jian, Z., Lv, W., Chen, Y., Luo, G., Qiu, W., Wang, S., Hu, B., Wu, Q.: MDF: A modality-aware disentanglement and fusion framework for multimodal sentiment analysis. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 40, pp. 31292–31300 (2026). <https://doi.org/10.1609/aaai.v40i37.40392>
- Kornblith, S., Norouzi, M., Lee, H., Hinton, G.: Similarity of neural network representations revisited. In: Proceedings of the 36th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 97, pp. 3519–3529 (2019)
- Liu, Z., Cai, L., Yang, W., Liu, J.: Sentiment analysis based on text information enhancement and multimodal feature fusion. Pattern Recognition **156**, 110847 (2024) <https://doi.org/10.1016/j.patcog.2024.110847>
- Lin, T.-Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2999–3007 (2017). <https://doi.org/10.1109/ICCV.2017.324>
- Li, Z., Huang, Z., He, X., Yu, J., Chen, H., Yang, C., Pan, Y.: Representation distribution matching and dynamic routing interaction for multimodal sentiment analysis. Knowledge-Based Systems **316**, 113376 (2025) <https://doi.org/10.1016/j.knosys.2025.113376>
- Liu, Z., Shen, Y., Lakshminarasimhan, V.B., Liang, P.P., Zadeh, A.B., 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, pp. 2247–2256 (2018). <https://doi.org/10.18653/v1/P18-1209>
- Liu, Y., Wang, S.: FINE: Factorized multimodal sentiment analysis via mutual information estimation. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 40, pp. 1955–1963 (2026). <https://doi.org/10.1609/aaai.v40i3.37176>
- Lin, J., Wang, Y., Xu, Y., Liu, Q.: Semi-IIN: Semi-supervised intra-inter modal interaction learning network for multimodal sentiment analysis. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 39, pp. 1411–1419 (2025). <https://doi.org/10.1609/aaai.v39i2.32131>
- Li, X., Zhang, H., Dong, Z., Cheng, X., Liu, Y., Zhang, X.: Learning fine-grained representation with token-level alignment for multimodal sentiment analysis. Expert Systems with Applications **269**, 126274 (2025) <https://doi.org/10.1016/j.eswa.2024.126274>
- Ou, L., Li, Z.: TPFN: A text-guided progressive fusion network for multimodal sentiment analysis. IEEE Transactions on Affective Computing **17**(2), 1585–1597 (2026)

<https://doi.org/10.1109/TAFFC.2026.3663280>

- 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). <https://doi.org/10.18653/v1/2020.acl-main.214>
- Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J.: Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research* **21**(140), 1–67 (2020)
- Stöckli, S., Schulte-Mecklenbeck, M., Borer, S., Samson, A.C.: Facial expression analysis with AFFDEX and FACET: A validation study. *Behavior Research Methods* **50**(4), 1446–1460 (2018) <https://doi.org/10.3758/s13428-017-0996-1>
- 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, pp. 6558–6569 (2019). <https://doi.org/10.18653/v1/P19-1656>
- Tao, C., Li, J., Zang, T., Gao, P.: A multi-focus-driven multi-branch network for robust multimodal sentiment analysis. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 39, pp. 1547–1555 (2025). <https://doi.org/10.1609/aaai.v39i2.32146>
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: Advances in Neural Information Processing Systems, vol. 30, pp. 5998–6008 (2017)
- 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) <https://doi.org/10.1016/j.patcog.2022.109259>
- Wu, S., He, D., Wang, X., Wang, L., Dang, J.: Enriching multimodal sentiment analysis through textual emotional descriptions of visual-audio content. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 39, pp. 1601–1609 (2025). <https://doi.org/10.1609/aaai.v39i2.32152>
- Wang, D., Liu, S., Wang, Q., Tian, Y., He, L., Gao, X.: Cross-modal enhancement network for multimodal sentiment analysis. *IEEE Transactions on Multimedia* **25**, 4909–4921 (2023) <https://doi.org/10.1109/TMM.2022.3183830>
- Wang, Y., Shen, Y., Liu, Z., Liang, P.P., Zadeh, A., Morency, L.-P.: Words can shift: Dynamically adjusting word representations using nonverbal behaviors. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, pp. 7216–7223 (2019). <https://doi.org/10.1609/aaai.v33i01.33017216>

- Wang, C., Zhang, W., Chen, G., Dong, M., He, T., Jiang, X.: Divide-and-conquer: Prompt-based distribution learning for multimodal sentiment analysis. *Information Fusion* **133**, 104293 (2026) <https://doi.org/10.1016/j.inffus.2026.104293>
- Wang, P., Zhou, Q., Wu, Y., Chen, T., Hu, J.: DLF: Disentangled-language-focused multimodal sentiment analysis. In: *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, pp. 21180–21188 (2025). <https://doi.org/10.1609/aaai.v39i20.35416>
- Xu, Q., Wei, Y., Wu, C., Wang, L., Yuan, S., Wu, J., Lu, J., Zhou, H.: Towards multimodal sentiment analysis via hierarchical correlation modeling with semantic distribution constraints. In: *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, pp. 21788–21796 (2025). <https://doi.org/10.1609/aaai.v39i20.35484>
- Yang, Y., Dong, X., Qiang, Y.: CLGSI: A multimodal sentiment analysis framework based on contrastive learning guided by sentiment intensity. In: *Findings of the Association for Computational Linguistics: NAACL 2024*, pp. 2099–2110 (2024). <https://doi.org/10.18653/v1/2024.findings-naacl.135>
- Yang, Y., Dong, X., Qiang, Y.: MSE-Adapter: A lightweight plugin endowing LLMs with the capability to perform multimodal sentiment analysis and emotion recognition. In: *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, pp. 25642–25650 (2025). <https://doi.org/10.1609/aaai.v39i24.34755>
- Yang, K., Xu, H., Gao, K.: CM-BERT: Cross-modal BERT for text-audio sentiment analysis. In: *Proceedings of the 28th ACM International Conference on Multimedia*, pp. 521–528 (2020). <https://doi.org/10.1145/3394171.3413690>
- 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). <https://doi.org/10.1609/aaai.v35i12.17289>
- 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*, pp. 1103–1114 (2017). <https://doi.org/10.18653/v1/D17-1115>
- Zhang, W., Du, Y., Wang, H., Liu, J., Li, X.: SSEDf: A shared-private semantic enhanced dynamic fusion network for multimodal sentiment analysis. *Expert Systems with Applications* **303**, 130667 (2026) <https://doi.org/10.1016/j.eswa.2025.130667>
- Zhu, A., Hu, M., Wang, X., Yang, J., Tang, Y., Ren, F.: KEBR: Knowledge enhanced self-supervised balanced representation for multimodal sentiment analysis. In: *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 5732–5741

(2024). <https://doi.org/10.1145/3664647.3681163>

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). <https://doi.org/10.1609/aaai.v32i1.12021>

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, pp. 2236–2246 (2018). <https://doi.org/10.18653/v1/P18-1208>

Zhi, Y., Li, J., Wang, H., Chen, J., Cao, T.: A fine-grained tri-modal interaction model for multimodal sentiment analysis. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 5715–5719 (2024). <https://doi.org/10.1109/ICASSP48485.2024.10447872>

Zhang, H., Wang, Y., Yin, G., Liu, K., Liu, Y., Yu, T.: Learning language-guided adaptive hyper-modality representation for multimodal sentiment analysis. In: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pp. 756–767 (2023)

Zhao, X., Yang, S., Tang, B.: CCAF: Coarse-to-fine cross-modal alignment and fusion for multimodal sentiment analysis. In: Proceedings of the ACM Web Conference 2026, pp. 7421–7430. ACM, New York (2026). <https://doi.org/10.1145/3774904.3792569>

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) <https://doi.org/10.1109/MIS.2016.94>