Computers, Materials & Continua

Tech Science Press

https://doi.org/10.32604/cmc.2026.087279

ARTICLE

DMHG-LEDS: Joint Differentiated Modality-Aware Heterogeneous Graph and Local Emotion Difference Supervision for Multimodal Emotion Recognition in Conversations

Yu Chen[1], Panpan Chen[1], Jun Wu[1,2,3], Shuai Guo[1], Jiahui Huang[1], Xinyi Zhu[1] and Qun Zhang[1,*]

[1]School of Computer Science and Artificial Intelligence, Hubei University of Technology, Wuhan, China
[2]Hubei Provincial Key Laboratory of Green Intelligent Computing Power Network, Hubei University of Technology, Wuhan, China
[3]Hubei Provincial Engineering Research Center for Digital & Intelligent Manufacturing Technologies and Applications, Hubei University of Technology, Wuhan, China

*Corresponding Author: Qun Zhang. Email: zhang.qun.hbut@gmail.com

Received: 15 June 2026; Accepted: 24 July 2026; Published: 15 September 2026

ABSTRACT: Multimodal Emotion Recognition in Conversations (MERC) has garnered substantial research attention recently. Existing MERC methods face several challenges: (1) they apply shared or coarse-grained graph construction rules across modalities, overlooking their distinct dependency patterns; (2) they rely on fixed-activation MLPs for feature transformation, limiting nonlinear representation capacity in complex emotional scenarios; (3) they focus predominantly on contextual modeling while underexploring local emotion discrimination between related utterances. To address these issues, we propose Joint Differentiated Modality-Aware Heterogeneous Graph and Local Emotion Difference Supervision for Multimodal Emotion Recognition in Conversations (DMHG-LEDS), a novel MERC framework. Specifically, modality-aware heterogeneous graphs are constructed by assigning differentiated intra-modal connection strategies to text, visual, and audio modalities, enabling modality-dependent contextual relation modeling. Based on the resulting graph topology, graph convolution aggregates neighborhood information and Chebyshev-KAN subsequently performs adaptive nonlinear transformation within each propagation step. Finally, an Entropy-Gated Local Emotion Difference Supervision module is introduced as an auxiliary task. It constructs ordered utterance pairs within non-overlapping local segments and provides entropy-gated supervision based on emotion-label differences, thereby improving the discriminability of local utterance representations. Experiments on Interactive emotional dyadic motion capture database (IEMOCAP), A Multimodal Multi-Party Dataset for Emotion Recognition in Conversation (MELD), and Multimodal Language Analysis in the Wild: CMU-MOSEI Dataset and Interpretable Dynamic Fusion Graph (CMU-MOSEI) demonstrate the effectiveness of the proposed method, which outperforms all baselines.

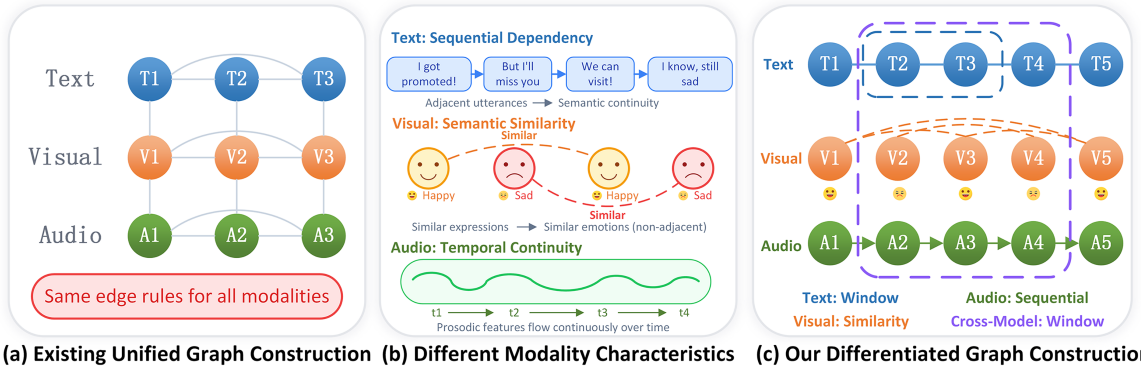
KEYWORDS: Multimodal emotion recognition; local emotion difference supervision; graph neural networks; Kolmogorov-Arnold Networks; multimodal fusion

1 Introduction

Multimodal Emotion Recognition in Conversation (MERC) aims to accurately identify the emotional state of each utterance by comprehensively utilizing multi-source information such as text, visual, and audio cues from continuous dialogue contexts [1,2]. With the widespread application of intelligent dialogue systems in scenarios such as social interaction, human-computer collaboration, and mental health monitoring [3,4], dialogue systems with robust emotion understanding capabilities have become an important research objective in the field of affective computing. However, in real dialogue scenarios, emotions are often expressed collaboratively through multiple modalities and evolve dynamically with the progression of the

conversation, which poses significant challenges [5–8] for the MERC task in terms of structural modeling and emotion dynamics modeling.

In recent years, Graph Neural Networks (GNNs) have been widely introduced into MERC tasks due to their advantages in modeling structural relationships [9,10]. Representative works [11–14] construct dialogue graphs to explicitly model utterances and their cross-speaker and cross-modal relationships, thereby effectively capturing contextual dependencies. Although graph-based methods have achieved significant progress, several key issues remain. First, in multimodal conversational emotion recognition, each modality involves multiple types of dependency relations when conveying emotional information; however, the dominant dependency patterns differ significantly across modalities (as illustrated in Fig. 1b). The text modality expresses emotions primarily through word semantics, contextual coherence, and logical transitions between utterances, so its emotional dependencies typically manifest as local sequential relations. The visual modality reflects emotional states mainly via facial expressions, eye gaze, and postural changes; since similar expression patterns may appear in non-adjacent utterances, its dependencies are better characterized as semantic similarity across time. The audio modality conveys emotions through variations in pitch, energy, speech rate, and prosody, exhibiting stronger local temporal continuity. Many existing graph-based MERC methods employ shared or coarse-grained intra-modal connection rules (as shown in Fig. 1a) [5,15]. Such shared rules may introduce irrelevant neighbors or overlook informative modality-specific relations. In contrast, cross-modal edges encode temporal correspondence across modalities and can retain a unified alignment rule. Second, existing graph neural network methods generally rely on multi-layer perceptrons (MLPs) with fixed activation functions for nonlinear transformation during the node feature update stage [9,16]. However, multimodal conversational emotion involves multiple nonlinear phenomena, including nonlinear combinatorial effects across modalities, nonlinear context-dependent relationships between emotion labels and utterance features, and nonlinear responses of emotion intensity to feature variation. These nonlinear relationships manifest in diverse forms across different nodes and conversational scenarios. Under such circumstances, fixed-activation transformations may be insufficient to adaptively reconstruct heterogeneous contextual representations under different dialogue contexts. Third, existing models often rely heavily on contextual information for prediction, while local emotion-difference relationships between nearby utterances are insufficiently explored. Constructing pairwise samples over broad dialogue ranges may introduce redundant long-range comparisons, while assigning equal importance to all pairs may overlook their different representation uncertainties [17–19]. Therefore, it is necessary to jointly model modality-specific structural dependencies, adaptive nonlinear contextual transformation, and local emotion discrimination in MERC.



(a) Existing Unified Graph Construction (b) Different Modality Characteristics (c) Our Differentiated Graph Construction

Figure 1: Illustrates two distinct multimodal graph construction approaches. (**a**) presents an existing unified graph construction method. (**b**) provides a detailed analysis of the unique properties of each modality through their respective feature components. (**c**) shows our purposefully designed edge construction strategy for the graph.

To address the above challenges, this paper proposes the Joint Differentiated Modality-Aware Heterogeneous Graph and Local Emotion Difference Supervision for Multimodal Emotion Recognition in Conversations (DMHG-LEDS) model. In the graph construction stage, we design differentiated intra-modal connection strategies according to the dominant dependency pattern of each modality, as shown in Fig. 1c: a position-window strategy is used for text to capture local semantic continuity, a similarity-driven strategy is used for vision to discover cross-temporal expression similarity, and a sequential adjacency strategy is adopted for audio to preserve local prosodic continuity. Cross-modal edges retain a uniform window-based alignment rule, since temporal co-occurrence between modalities is a structural alignment relation rather than a mechanism-specific one. In the graph propagation stage, we design a propagation network with alternately stacked graph convolution and Chebyshev Polynomial-Based Kolmogorov-Arnold Networks: An Efficient Architecture for Nonlinear Function Approximation (Chebyshev-KAN) [20] layers, where graph convolution first aggregates heterogeneous contextual information from the selected neighborhood, and Chebyshev Polynomial-Based Kolmogorov-Arnold Networks: An Efficient Architecture for Nonlinear Function Approximation (Chebyshev-KAN) subsequently performs adaptive nonlinear transformation on the aggregated representation. In addition, we introduce an Entropy-Gated Local Emotion Difference Supervision module as an auxiliary task. It constructs ordered utterance pairs within non-overlapping local segments and provides entropy-gated supervision according to whether their emotion labels differ, thereby improving the discriminability of local utterance representations Extensive experiments on Interactive emotional dyadic motion capture database (IEMOCAP), A Multimodal Multi-Party Dataset for Emotion Recognition in Conversation (MELD), and Multimodal Language Analysis in the Wild: CMU-MOSEI Dataset and Interpretable Dynamic Fusion Graph (CMU-MOSEI) demonstrate the effectiveness of the proposed method, with weighted F1 scores of 73.29%, 66.55%, and 46.54%, respectively. Comprehensive ablation studies further verify the contribution of each component. The main contributions of this paper are as follows:

(1) We propose a differentiated modality-aware heterogeneous graph for Multimodal Emotion Recognition in Conversation. Instead of applying a uniform intra-modal connection rule, the proposed graph explicitly models local sequential dependencies for text, similarity-based dependencies for vision, and temporal continuity for audio, thereby providing modality-specific contextual neighborhoods for graph propagation.
(2) We design a Chebyshev Polynomial-Based Kolmogorov-Arnold Networks: An Efficient Architecture for Nonlinear Function Approximation (Chebyshev-KAN) based alternating graph propagation mechanism. Graph convolution first aggregates contextual information according to the differentiated graph topology, and Chebyshev-KAN then adaptively transforms the aggregated heterogeneous representation, enhancing nonlinear contextual modeling within each propagation step.
(3) We introduce an Entropy-Gated Local Emotion Difference Supervision module. By constructing ordered utterance pairs within non-overlapping local segments and adaptively weighting uncertain pairs, the module provides auxiliary supervision to improve the discriminability of local utterance representations for emotion classification.

2 Related Work

2.1 Multimodal Affective Computing

Multimodal affective computing includes tasks such as multimodal sentiment analysis and emotion recognition. It integrates textual, visual, and acoustic information to identify an individual's emotional state or sentiment tendency. Early studies mainly employed feature fusion, attention mechanisms, and cross-modal interaction to integrate information from multiple sources [7,8]. Contextual interaction-based

multimodal emotion analysis with enhanced semantic information (CIME) further improves multimodal emotion representation through contextual interaction and semantic enhancement [21]. With the development of this field, researchers have increasingly focused on the differences in contributions among modalities. Modalities are not always consistent in terms of expressive capability, data quality, and affective cues, which may affect fusion performance. To address this issue, uncertainty-aware methods estimate modality reliability and adapt the fusion process accordingly [22]. Under noisy or incomplete input conditions, robust representation learning and feature restoration methods can alleviate the influence of missing information [23]. Prompt learning also provides a new modeling paradigm for sentiment analysis and emotion recognition with missing modalities [24]. In addition, pre-trained models and distributed learning have further expanded the application scope of multimodal affective computing. Few-shot multimodal sentiment analysis for social media via integrated prompt learning and vision-language models (FMSA) combines integrated prompt learning with vision-language models for few-shot multimodal sentiment analysis in social media [25]. Federated multimodal affective computing supports distributed training under privacy constraints [26]. These studies have advanced multimodal affective computing from the perspectives of semantic interaction, fusion reliability, robustness, and learning paradigms. For continuous dialogue scenarios, MERC needs to jointly model inter-utterance context, speaker interaction, and cross-modal information. How to effectively capture these relationships has therefore become a central concern of subsequent MERC studies.

2.2 Graph Neural Networks for ERC

Graph neural networks have been widely applied to conversational emotion recognition tasks due to their ability to model structured relationships. A graph convolutional neural network for emotion recognition in conversation (DialogueGCN) [16] first introduced graph convolutional networks to this task, constructing dialogue graphs based on speaker relationships and utterance positions. Directed acyclic graph network for conversational emotion recognition (DAG-ERC) [27] employed directed acyclic graphs to model the temporal dependencies in conversations. In multimodal scenarios, Multimodal fusion via deep graph convolution network for emotion recognition in conversation (MMGCN) [9] treated utterances from different modalities as graph nodes, achieving cross-modal information interaction through graph convolution. Multimodal dynamic fusion network for emotion recognition in conversations (MM-DFN) [10] introduced a dynamic fusion mechanism to adaptively adjust modality weights. A directed graph based cross-modal feature complementation approach for multimodal conversational emotion recognition (GraphCFC) [11] modeled cross-modal feature complementarity relationships through directed graphs. Multivariate, multi-frequency and multimodal: Rethinking graph neural networks for emotion recognition in conversation (M3Net) [12] designed a multi-frequency signal decomposition module from a frequency domain perspective to process multimodal features. These methods have demonstrated the effectiveness of graph structures for multimodal conversational emotion recognition. However, existing graph-based approaches mainly focus on contextual propagation, dynamic fusion, or cross-modal interaction, while the modality-dependent characteristics of intra-modal relations are not always explicitly encoded in the graph topology: text exhibits sequential dependencies, with adjacent utterances having semantic continuity; visual information reflects semantic similarity more, where similar facial expressions often correspond to similar emotional states; Audio signals exhibit temporal continuity, with prosodic features of adjacent utterances maintaining a sequential relationship. To address this issue, our method explicitly assigns differentiated intra-modal connection rules to text, visual, and audio modalities within a unified heterogeneous graph, thereby providing modality-specific contextual neighborhoods before graph propagation.

2.3 Kolmogorov-Arnold Networks

Kolmogorov-Arnold Networks (KAN) [28] are a novel network architecture based on the Kolmogorov-Arnold representation theorem. Unlike MLPs that apply fixed activation functions at nodes, KAN places learnable univariate functions on network edges, approximating complex multivariate functions through linear combinations of nonlinear functions. The original KAN utilizes B-spline functions and demonstrates superior parameter efficiency, but its complex grid computation leads to high resource overhead. To address this, researchers have proposed various improved variants: Chebyshev Polynomial-Based Kolmogorov-Arnold Networks: An Efficient Architecture for Nonlinear Function Approximation (Chebyshev-KAN) [20] leverages the orthogonal properties of Chebyshev polynomials to significantly reduce computational complexity. Wavelet Kolmogorov-Arnold Networks (Wav-KAN) [29] introduces wavelet functions as basis functions, effectively capturing local time-frequency features in data through the multi-resolution analysis advantage of wavelet transforms. Implicit neural representations with fourier kolmogorov-arnold networks (FourierKAN) [30] employs Fourier series parameterization, efficiently modeling periodic patterns and high-frequency components in data through frequency domain transformation mechanisms. Although KAN has been applied to graph learning, time-series prediction, and image processing, its use in MERC remains limited. Existing graph-based MERC methods commonly adopt fixed-activation MLPs for post-aggregation feature transformation. Different from using KAN as an independent feature enhancement module, our method places Chebyshev-KAN after each graph convolution step, so that the contextual information aggregated from modality-specific and cross-modal neighborhoods can be adaptively reconstructed through learnable polynomial-based nonlinear transformation within the graph propagation process.

2.4 Local Emotion Difference Supervision

Emotional states in conversations evolve dynamically, and modeling local affective differences is beneficial for emotion recognition. Previous studies have introduced emotion- or sentiment-related auxiliary tasks to improve representation learning. Emotion Recognition in Conversations (ERC) with emotion shift detection based on multi-task learning (ERC-ESD) [17] jointly modeled emotion recognition and affective state changes, while A Cross-Modal Fusion Network with Emotion-Shift Awareness for Dialogue Emotion Recognition (CFN-ESA) [31] introduced an affective change awareness mechanism to capture multimodal cues of emotional changes. These studies demonstrate that auxiliary supervision based on affective differences can benefit ERC. However, existing auxiliary objectives may construct utterance pairs over broad dialogue ranges or assign equal importance to all pairwise samples, which can introduce redundant long-range comparisons and overlook the varying uncertainty of local emotion relationships. To address this issue, the proposed Entropy-Gated Local Emotion Difference Supervision (LEDS) module restricts pair construction to non-overlapping local segments and employs entropy-based gating to adaptively adjust the contribution of local utterance pairs.

3 Method

3.1 Framework Overview

Given a conversation of M utterances, where each utterance u_i $(i = 1, \ldots, M)$ contains textual (t), visual (v), and audio (a) modalities, the MERC task aims to predict the emotion label e_i for each u_i. The overall architecture of Joint Differentiated Modality-Aware Heterogeneous Graph and Local Emotion Difference Supervision for Multimodal Emotion Recognition in Conversations (DMHG-LEDS) is illustrated in Fig. 2, comprising utterance-level feature extraction, differentiated modality-aware heterogeneous graph construction, Chebyshev-KAN-based alternating graph propagation, entropy-gated local emotion difference supervision, and emotion classification.

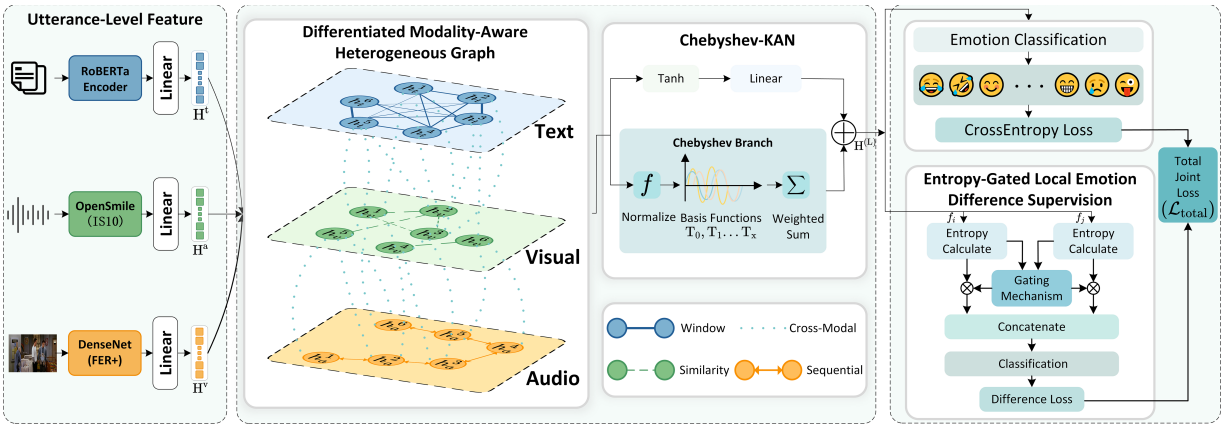


Figure 2: Overall architecture of DMHG-LEDS.

3.2 Utterance-Level Feature Extraction

Text features $x_i^t \in \mathbb{R}^{d_t}$ are extracted using the pre-trained Robustly Optimized BERT Pretraining Approach (RoBERTa) model via the `[CLS]` token. Visual features $x_i^v \in \mathbb{R}^{d_v}$ are obtained from DenseNet pre-trained on the Facial Expression Recognition Plus (FER+) dataset. Audio features $x_i^a \in \mathbb{R}^{d_a}$ are extracted using the INTERSPEECH 2010 Paralinguistic Challenge feature set (IS10) configuration of the OpenSmile toolkit. To align all modalities into a shared d-dimensional space, three independent projection layers are applied:

$$h_i^m = \text{Dropout}(\text{LeakyReLU}(x_i^m W^m + b^m)), \quad m \in \{t, v, a\}, \tag{1}$$

where $W^m \in \mathbb{R}^{d_m \times d}$ and $b^m \in \mathbb{R}^d$ are learnable parameters, and d is the unified hidden dimension.

3.3 Multimodal Heterogeneous Conversational Graph

This section details the multimodal heterogeneous conversational graph construction module. The module comprises two core components: a differentiated modality-aware heterogeneous graph that designs dedicated edge connection rules for each modality (Section 3.3.1), and a Chebyshev-KAN based alternating propagation network that achieves efficient feature aggregation and nonlinear updates (Section 3.3.2).

3.3.1 Differentiated Modality-Aware Heterogeneous Graph

We construct a unified tri-modal heterogeneous dialogue graph $G = (V, E, W)$ with differentiated modality-aware construction strategies, enabling the model to accurately capture both intra- and inter-modal affective dependencies.

Node Definition: The node set $V = \{V^t \cup V^v \cup V^a\}$ comprises nodes from three modalities. For a dialogue with M utterances, the graph contains $3M$ nodes in total: $V^t = \{v_1^t, \ldots, v_M^t\}$, $V^v = \{v_1^v, \ldots, v_M^v\}$, and $V^a = \{v_1^a, \ldots, v_M^a\}$. Each node v_i^m carries feature vector $h_i^m \in \mathbb{R}^d$, where $m \in \{t, v, a\}$ and d is the unified hidden dimension.

Edge Set Definition: The edge set $E = E^t \cup E^v \cup E^a \cup E^{\text{cross}}$ defines node connectivity. Since graph convolution aggregates information from the neighbors specified by the graph topology, the intra-modal edge rule determines the contextual evidence directly available to each node. We therefore use different connection rules for the three modalities: a position-window graph for text to preserve local discourse context, a similarity-driven graph for vision to connect visually related utterances across distant dialogue

positions, and a sequential graph for audio to model short-range prosodic evolution. The corresponding edge sets are defined as follows.

Position-Window Strategy for Text Modality: Text information exhibits strong sequential dependencies, where adjacent utterances often have semantic or emotional connections. We adopt a fixed-window strategy: for any text node v_i^t, bidirectional edges are established with nodes within window w_{text}:

$$E^t = \{(v_i^t, v_j^t), (v_j^t, v_i^t) \mid |i-j| \le w_{\text{text}}\}, \tag{2}$$

where w_{text} is selected on the validation set. This construction restricts the direct neighborhood of each text node to nearby dialogue turns, allowing graph propagation to focus on local discourse context.

Similarity-Driven Strategy for Visual Modality: Unlike text, visual information (e.g., facial expressions) exhibits semantic-level correlations, where similar expressions may correspond to similar emotional states even across non-adjacent utterances. We adopt a dynamic construction method based on feature similarity. Visual features are L2-normalized as $\widetilde{h}_i^v = h_i^v / \|h_i^v\|_2$, and cosine similarity is computed:

$$\text{sim}(v_i^v, v_j^v) = \tilde{h}_i^v \cdot \tilde{h}_j^v, \tag{3}$$

Edges are established only when similarity exceeds threshold τ_{sim}:

$$E^v = \{(v_i^v, v_j^v) \mid \text{sim}(v_i^v, v_j^v) > \tau_{\text{sim}},\ i \ne j\}, \tag{4}$$

where τ_{sim} is selected on the validation set. This construction enables visually related utterances to exchange information directly even when they occur at distant positions in the dialogue.

Sequential Adjacency Strategy for Audio Modality: Audio signals possess natural temporal continuity, where the prosody of adjacent utterances often exhibits continuation or contrast patterns. We establish connections only between each audio node v_i^a and its immediate neighbor:

$$E^a = \{(v_i^a, v_{i+1}^a) \mid v_i^a, v_{i+1}^a \in V^a\}, \tag{5}$$

This construction restricts direct audio aggregation to adjacent dialogue turns, thereby preserving short-range prosodic continuity during graph propagation.

Cross-Modal Connection Construction: To enable inter-modal information interaction, we apply a unified window-based strategy. For a modality m_1 node $v_i^{m_1}$, connections are established with all modality m_2 nodes within window $[i - w_p, i + w_f]$:

$$E^{\text{cross}} = \{(v_i^{m_1}, v_j^{m_2}) \mid m_1 \ne m_2,\ i - w_p \le j \le i + w_f,\ m_1, m_2 \in \{t, v, a\}\}, \tag{6}$$

The complete edge set is thus:

$$E = E^t \cup E^v \cup E^a \cup E^{\text{cross}}, \tag{7}$$

Accordingly, the proposed graph establishes three distinct intra-modal neighborhoods before graph propagation: a local temporal neighborhood for text, a similarity-based non-local neighborhood for vision, and a sequential neighborhood for audio. Cross-modal edges further connect temporally aligned nodes across modalities. In this way, the heterogeneous graph jointly encodes modality-specific contextual relations and inter-modal temporal alignment.

3.3.2 Chebyshev-KAN Based Alternating Graph Propagation

After constructing the heterogeneous dialogue graph $G = (V, E, W)$, graph neural networks propagate and update node features according to the modality-specific and cross-modal relations encoded in the graph topology. However, the aggregated representations may contain complex nonlinear interactions among modality-dependent contextual cues, cross-modal alignment information, and dialogue states. Standard MLPs with fixed activation functions are insufficient for this adaptive nonlinear modeling. Therefore, we introduce a Chebyshev-KAN layer after each graph convolution operation. In each propagation step, graph convolution first aggregates neighborhood information selected by the differentiated topology, and Chebyshev-KAN subsequently performs adaptive nonlinear transformation on the aggregated representation.

Graph Convolution Neighborhood Aggregation: An adjacency matrix $A \in \mathbb{R}^{3M\times 3M}$ is constructed from the edge set, where $A_{ij} = w_{ij}$ if $(v_i, v_j) \in E$ and $A_{ij} = 0$ otherwise, with each w_{ij} initialized to 1. Self-loops are added via $\hat{\mathbf{A}} = \mathbf{A} + \mathbf{I}$. Let $\hat{\mathbf{D}}$ denote the corresponding structural degree matrix, where $\hat{D}_{ii} = \sum_j \mathbb{I}[\hat{A}_{ij} \neq 0]$. The normalized adjacency matrix is defined as:

$$\tilde{A} = \hat{D}^{-\frac{1}{2}}\hat{A}\hat{D}^{-\frac{1}{2}}, \tag{8}$$

Since a self-loop is added to every node, $\hat{D}_{ii} \geq 1$ holds for all nodes. The graph convolution for layer L is defined as:

$$H^{(L)}_{\text{gcn}} = \tilde{A}H^{(L-1)}W^{(L)}_{\text{gcn}}, \tag{9}$$

where $W^{(L)}_{\text{gcn}} \in \mathbb{R}^{d\times d}$ is a learnable weight matrix. Through this operation, each node aggregates contextual information from its graph neighborhood.

Chebyshev-KAN Non-linear Transformation: Following graph convolution, a Chebyshev-KAN layer performs non-linear feature transformation via learnable Chebyshev polynomial basis functions. Input features are first mapped to a bounded interval:

$$x_{\text{norm}} = \tanh(\text{clamp}(x, -1, 1)), \tag{10}$$

Then Chebyshev polynomial basis functions are computed:

$$T_n(x_{\text{norm}}) = \cos(n \cdot \arccos(x_{\text{norm}})), \quad n = 0, 1, \ldots, K, \tag{11}$$

The final transformation integrates a base linear component with the polynomial expansion:

$$\text{Chebyshev-KAN}(x) = W_{\text{base}} \cdot \sigma(x) + \sum_{n=0}^{K} c_n T_n(x_{\text{norm}}) + b, \tag{12}$$

where W_{base}, c_n, and b are learnable parameters, and $\sigma(\cdot)$ is instantiated as the Tanh function in our implementation. The maximum polynomial order is set to $K = 7$. For a Chebyshev-KAN layer with input and output dimensions both equal to d, the parameter number is $(K + 2)d^2 + d$, and the computational complexity is approximately $\mathcal{O}((K + 2)d^2)$. Although this introduces additional polynomial-basis computation compared with a standard MLP, the fixed low-order setting keeps the overhead manageable.

Alternating Propagation Architecture: The propagation for each layer proceeds as follows: neighborhood information is first aggregated via graph convolution; the Chebyshev-KAN layer then applies

non-linear transformation with Dropout regularization; finally, a residual connection preserves multi-layer features:

$$H^{(L)} = H^{(L-1)} + \text{Dropout}(\text{Chebyshev-KAN}(H_{\text{gcn}}^{(L)})), \tag{13}$$

where the initial feature matrix $H^{(0)} = [h_1^t, \ldots, h_M^t, h_1^v, \ldots, h_M^v, h_1^a, \ldots, h_M^a]^T$ is obtained from Eq. (1). After L propagation layers, the final node representations $H^{(L)} \in \mathbb{R}^{3M \times d}$ are obtained.

3.4 Entropy-Gated Local Emotion Difference Supervision

To improve the discriminability of utterance-level emotion representations, we introduce an Entropy-Gated Local Emotion Difference Supervision (LEDS) module as an auxiliary task. Specifically, each dialogue is partitioned into non-overlapping local segments, and all ordered utterance pairs are constructed within each segment. An entropy-based gating mechanism then adaptively weights the pairs according to their representation uncertainty, while binary supervision is provided based on whether the two utterances share the same emotion category. This auxiliary task provides additional local supervision for distinguishing utterance pairs with the same and different emotion labels.

3.4.1 Pairwise Emotion-Difference Feature Construction

After L layers of heterogeneous graph propagation, the feature matrix $H^{(L)} \in \mathbb{R}^{3M \times d}$ is obtained. For the i-th utterance, the node features across text, visual, and audio modalities are $H_i^{(L)}$, $H_{M+i}^{(L)}$, and $H_{2M+i}^{(L)}$, respectively. The utterance-level fused representation is obtained by concatenating and projecting the tri-modal features:

$$f_i = \text{LeakyReLU}(W_f[H_i^{(L)}; H_{M+i}^{(L)}; H_{2M+i}^{(L)}] + b_f), \tag{14}$$

where $W_f \in \mathbb{R}^{3d \times d}$ and $b_f \in \mathbb{R}^d$ are learnable parameters. This yields the fused feature sequence $F = \{f_1, f_2, \ldots, f_M\}$.

To avoid noisy supervision from distant utterance pairs and control the computational cost, we divide each dialogue into consecutive non-overlapping local segments of length ω. When $\omega > 0$, all ordered utterance pairs are enumerated within each segment rather than randomly sampled. If the r-th segment contains s_r utterances, where $s_r = \min \omega, M - (r-1)\omega$, it produces s_r^2 pairwise samples. Accordingly, the total number of pairs for a dialogue is $N_{\text{pairs}} = \sum_{r=1}^{\lceil M/\omega \rceil} s_r^2$. When $\omega = -1$, no local segmentation is performed, and all ordered utterance pairs are constructed over the entire dialogue, yielding M^2 samples. All ordered pairs within each local segment are retained, including self-pairs (u_i, u_i) and bidirectional pairs (u_i, u_j) and (u_j, u_i). The binary emotion-difference label for each pair is assigned according to Eq. (20), and pairwise supervision is generated only for utterances within the same local segment. Compared with global pairwise construction with complexity $\mathcal{O}(M^2 d)$, the proposed local segmentation strategy reduces the pairwise feature construction cost to approximately $\mathcal{O}(M\omega d)$, where d denotes the utterance representation dimension. Therefore, ω controls both the supervision range and the training cost.

3.4.2 Entropy-Based Gating Mechanism

To softly reweight local utterance pairs according to their representation uncertainty, we introduce an entropy-based gating mechanism. For each utterance, the fused feature f_i is normalized via softmax and its Shannon entropy is computed as:

$$p_{i,k} = \frac{\exp(f_{i,k})}{\sum_{m=1}^{d} \exp(f_{i,m})}, \tag{15}$$

$$H(f_i) = -\sum_{k=1}^{d} p_{i,k} \log(p_{i,k} + \varepsilon), \tag{16}$$

where d is the feature dimension and $\varepsilon = 10^{-10}$ prevents numerical underflow. Higher entropy reflects greater distributional uncertainty, indicating more ambiguous emotional cues. The joint gated weight for utterance pair (f_i, f_j) is then computed as:

$$\alpha_{ij} = \sigma\left(\tanh\left(\frac{H(f_i) + H(f_j)}{2}\right)\right), \tag{17}$$

where $\sigma(\cdot)$ denotes the sigmoid function. The Tanh–Sigmoid mapping acts as a bounded soft gate, limiting the influence of extremely high-entropy pairs while preserving the relative uncertainty ordering among local pairs. Since the feature dimension is fixed within each dataset-specific model and the gate is only used for within-model pairwise reweighting, normalization by $\log d$ is not required. The gated pairwise features are constructed as:

$$f_i' = \alpha_{ij} \cdot f_i, \quad f_j' = \alpha_{ij} \cdot f_j, \tag{18}$$

The weighted features are concatenated and fed into a linear classifier for binary emotion-difference prediction:

$$z_{ij} = \text{softmax}(W_{\text{diff}}[f_i'; f_j'] + b_{\text{diff}}), \tag{19}$$

where $W_{\text{diff}} \in \mathbb{R}^{2d \times 2}$ and $b_{\text{diff}} \in \mathbb{R}^2$ are learnable parameters. Training labels are generated by comparing emotion categories of utterance pairs:

$$d_{ij} = \begin{cases} 1, & \text{if } r_i \neq r_j, \\ 0, & \text{if } r_i = r_j, \end{cases} \tag{20}$$

where r_i and r_j denote the emotion labels of u_i and u_j. The supervision loss is optimized via cross-entropy:

$$\mathcal{L}_{\text{diff}} = -\frac{1}{N_{\text{pairs}}} \sum_{k=1}^{N_{\text{pairs}}} \sum_{j=1}^{2} d_{k,j} \log z_{k,j}, \tag{21}$$

where N_{pairs} is the number of utterance pairs, $d_{k,j}$ is the one-hot ground truth label, and $z_{k,j}$ is the predicted probability. The loss $\mathcal{L}_{\text{diff}}$ serves as an auxiliary supervision signal during training. It regularizes the shared graph propagation and utterance-level fusion representations through local emotion-difference modeling, while the final emotion prediction is still produced by the classifier in Eq. (23). Since pairwise labels are derived from ground-truth emotion annotations, annotation inconsistencies may affect the auxiliary supervision for pairs involving mislabeled utterances. However, non-overlapping local segments confine this influence to local pairs, and $\mathcal{L}_{\text{diff}}$ is used only as an auxiliary training signal rather than for inference.

3.5 Emotion Classification

After L layers of heterogeneous graph propagation, the final node representation matrix $\mathbf{H}^{(L)} \in \mathbb{R}^{3M \times d}$ contains three modality-specific node representations for each utterance. For utterance u_i, the node indices

corresponding to the text, visual, and audio modalities are i, $M+i$, and $2M+i$, respectively. To obtain a unified utterance-level representation for emotion classification, we concatenate the three modalities node features and pass them through a linear fusion layer:

$$\mathbf{h}_i = \text{LeakyReLU}\left(\mathbf{W}_{\text{fuse}}\left[\mathbf{H}_i^{(L)}; \mathbf{H}_{M+i}^{(L)}; \mathbf{H}_{2M+i}^{(L)}\right] + \mathbf{b}_{\text{fuse}}\right), \tag{22}$$

where $\mathbf{W}_{\text{fuse}} \in \mathbb{R}^{3d\times d}$ and $\mathbf{b}_{\text{fuse}} \in \mathbb{R}^{d}$ are learnable parameters, and $[;]$ denotes the feature concatenation operation. This yields the utterance-level fused representation matrix $\mathbf{H} = \{\mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_M\} \in \mathbb{R}^{M\times d}$, which is then fed into a fully connected layer and a softmax layer for emotion category prediction:

$$\hat{\mathbf{Y}} = \text{softmax}(\mathbf{H}\,\boldsymbol{\theta}_y + \mathbf{b}_y), \tag{23}$$

$$e_i = \arg\max(\mathbf{y}_i), \tag{24}$$

where $\boldsymbol{\theta}_y$ and $\mathbf{b}_y$ are learnable parameters, $\mathbf{y}_i$ denotes the emotion probability vector of the i-th utterance u_i, and e_i denotes the corresponding predicted emotion label. In the training phase, we adopt the standard cross-entropy loss for emotion classification and the entropy-gated local emotion difference supervision loss to jointly optimize the network. The final loss function is formulated as follows:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{cls}} + \lambda_{\text{diff}}\mathcal{L}_{\text{diff}}, \tag{25}$$

where $\mathcal{L}_{\text{cls}} = -\frac{1}{M}\sum_{i=1}^{M}\sum_{j=1}^{C} y_{i,j}\log\hat{y}_{i,j}$ denotes the primary emotion classification loss, and $\mathcal{L}_{\text{diff}}$ denotes the auxiliary binary emotion-difference loss defined in Eq. (21). The coefficient λ_{diff} balances the contribution of local pairwise supervision and is selected according to the validation W-F1 score. The model is optimized using AdamW with a weight decay of 1×10^{-4}. During training, the classification loss and auxiliary difference loss jointly optimize the shared encoder; during inference, only the emotion classification branch is used for prediction.

4 Experiment

4.1 Datasets

To validate the effectiveness of the proposed model, we conduct experiments on three benchmark datasets: Interactive emotional dyadic motion capture database (IEMOCAP) [32], A Multimodal Multi-Party Dataset for Emotion Recognition in Conversation (MELD) [33], and Multimodal Language Analysis in the Wild: CMU-MOSEI Dataset and Interpretable Dynamic Fusion Graph (CMU-MOSEI) [34]. The statistical information is presented in Table 1. IEMOCAP [32] is a classic dyadic multimodal emotion dataset, comprising 151 dialogues and 7433 utterances, covering six emotion categories: happy, sad, neutral, angry, excited, and frustrated. Following the commonly used partition scheme [9,13], 120 dialogues are used as the training and validation set, and 31 dialogues as the test set, with 10% randomly sampled from the training data as the validation set. MELD [33] is a multi-party conversational emotion dataset sourced from the TV series *Friends*, comprising 1433 dialogues and 13,708 utterances, covering seven emotion categories: neutral, surprise, fear, sadness, joy, disgust, and anger. We adopt the predefined partition: 1153 dialogues as the training and validation set, and 280 dialogues as the test set. Furthermore, to comprehensively evaluate the generalisation capability of the model in real-world complex scenarios, we introduce the large-scale dataset CMU-MOSEI [34]. This dataset is collected from YouTube videos, comprising over 1000 speakers and 22,860 utterances. Given that its original labels are multi-label sparse annotations, following mainstream work [13], we employ it for seven-class sentiment analysis (ranging from highly negative to highly positive). The training and validation sets are merged, totalling 2549 dialogues (18,198 utterances), and the test set comprises 646 dialogues (4662 utterances), with 10% randomly sampled from the merged set as the validation set.

Table 1: Statistical information of all datasets.

Datasets	Dialogues			Utterances		
	Train	Valid	Test	Train	Valid	Test
IEMOCAP [32]	108	12	31	5163	647	1623
MELD [33]	1039	114	280	9989	1109	2610
CMU-MOSEI [34]	2294	255	646	16,378	1820	4662

Note: IEMOCAP, Interactive Emotional Dyadic Motion Capture Database; MELD, A Multimodal Multi-Party Dataset for Emotion Recognition in Conversation; CMU-MOSEI, Multimodal Language Analysis in the Wild: CMU-MOSEI Dataset and Interpretable Dynamic Fusion Graph.

4.2 Detailed Settings

Following MMGCN [9] and MM-DFN [10], Accuracy (ACC) and Weighted F1-score (W-F1) are adopted as the primary metrics. Results are reported as mean ± standard deviation over 10 random seeds, and paired t-tests against reproduced MM-DFN are conducted with $p < 0.05$. All experiments are implemented in PyTorch 2.4.0 with Compute Unified Device Architecture (CUDA) 12.5 on an NVIDIA GeForce RTX 4060Ti Graphics Processing Unit (GPU) and optimized using AdamW with a weight decay of 1×10^{-4}. The text window w_{text} was searched from 2 to 8 with a step of 1, τ_{sim} from 0.1 to 0.5 with a step of 0.1, and the symmetric cross-modal window size q, where $w_p = w_f = q$, from 1 to 40. The final settings were $w_{\text{text}} = 4$, $\tau_{\text{sim}} = 0.3$, and $(w_p, w_f) = (30, 30)$ for IEMOCAP, $(3, 3)$ for MELD, and $(5, 5)$ for CMU-MOSEI; ω was set to 19, 3, and 2 for IEMOCAP, MELD, and CMU-MOSEI, respectively. The Chebyshev coefficients were initialized from a zero-mean Gaussian distribution with standard deviation $0.1/\sqrt{d}$, and the dropout rate after Chebyshev-KAN was set to 0.2, 0.2, and 0.4 for IEMOCAP, MELD, and CMU-MOSEI, respectively, consistent with Table 2. The auxiliary loss coefficient λ_{diff} was set to 0.8 for IEMOCAP, 0.4 for MELD, and 1.0 for CMU-MOSEI.

Table 2: Partial hyperparameter description.

Dataset	Learning Rate	Batch Size	Dropout	Number of Network Layers	Cross-Modal Window	Training Epochs
IEMOCAP [32]	4×10^{-5}	16	0.2	6	30	150
MELD [33]	7×10^{-5}	16	0.2	5	3	50
CMU-MOSEI [34]	2×10^{-4}	32	0.4	2	5	100

Note: IEMOCAP, Interactive Emotional Dyadic Motion Capture Database; MELD, A Multimodal Multi-Party Dataset for Emotion Recognition in Conversation; CMU-MOSEI, Multimodal Language Analysis in the Wild: CMU-MOSEI Dataset and Interpretable Dynamic Fusion Graph.

4.3 Baseline Comparisons

To comprehensively evaluate the effectiveness of DMHG-LEDS, we compare our proposed model against multiple strong baselines. These baselines encompass classical multimodal fusion methods as well as advanced graph models incorporating specific mechanisms, such as reinforcement learning and dual emotion-semantic channels. Specifically, these include: Multimodal fusion via deep graph convolution network for emotion recognition in conversation (MMGCN) [9], Multimodal dynamic fusion network for emotion recognition in conversations (MM-DFN) [10], Dialog and Event Relation-Aware Graph Convolutional Neural Network for Multimodal Dialog Emotion Recognition (DER-GCN) [35], Learning More from

Mixed Emotions: A Label Refinement Method for Emotion Recognition in Conversations (EmoLR) [36], Multivariate, multi-frequency and multimodal: Rethinking graph neural networks for emotion recognition in conversation (M3Net) [12], GraphCFC: A directed graph based cross-modal feature complementation approach for multimodal conversational emotion recognition (GraphCFC) [11], Identity and Modality Attributes Driven Multimodal Fusion Networks for Emotion Recognition in Conversations (IMDNet) [37], Semantic and Emotional Dual Channel for Emotion Recognition in Conversation (SEDC) [38], RL-EMO: Reinforcement learning framework for multimodal emotion recognition (RL-EMO) [39], ECERC: Evidence-Cause Attention Network for Multi-Modal Emotion Recognition in Conversation (ECERC) [40], Multi-modal Anchor Gated Transformer with Knowledge Distillation for Emotion Recognition in Conversation (MAGTKD) [41], HiMul-LGG: A hierarchical decision fusion-based local-global graph neural network for multimodal emotion recognition in conversation (HiMul-LGG) [14], Supervised adversarial contrastive learning for emotion recognition in conversations (SACL-LSTM) [42], and DialogueCRN: Contextual reasoning networks for emotion recognition in conversations (DialogueCRN) [43].

4.4 Experimental Results and Analysis

Tables 3–5 report the performance comparison between our proposed model and baseline models on the IEMOCAP, MELD, and CMU-MOSEI datasets. Overall, our model achieves optimal performance on all three datasets. On the IEMOCAP dataset, our model achieves an accuracy of 73.20% and a W-F1 score of 73.29%. The W-F1 score shows improvements of 1.61% and 1.56% over recent state-of-the-art models SEDC and IMDNet, respectively. From the recognition performance across emotion categories, our model performs excellently on four emotion categories: Happy, Neutral, Excited, and Frustrated, with the W-F1 score for the Happy category reaching 61.64%, significantly outperforming the second-best model, HiMul-LGG.

On the MELD dataset, our proposed model achieves an accuracy of 68.12% and a W-F1 score of 66.55%, also superior to all baseline models. Although the improvement margin narrows compared to the IEMOCAP dataset, considering that the MELD dataset originates from dialogue clips from the TV series “Friends”, which contains substantial emotion-irrelevant background noise and more complex and variable dialogue scenarios, this improvement still holds significant practical value. Notably, our model achieves W-F1 scores of 28.87% and 59.58% on the minority emotion categories Fear and Surprise, respectively, significantly outperforming all baseline models. Furthermore, to verify the generalization capability of the model under different data scales and label settings, we conduct seven-class sentiment analysis experiments on CMU-MOSEI, with results shown in Table 5. DMHG-LEDS achieves 47.11% ACC and 46.54% W-F1, improving over MMGCN by 1.44 and 2.43 percentage points, respectively. CMU-MOSEI contains open-network video samples from more diverse sources and employs fine-grained sentiment tendency labels ranging from highly negative to highly positive, posing more challenging recognition conditions. Our method still achieves the best overall performance on this dataset, further validating its applicability and generalization across different multimodal emotion benchmarks. In summary, experimental results on the three datasets demonstrate that the proposed model can effectively model multimodal emotional information and maintain favorable recognition performance across different dialogue scenarios and emotion label settings.

Table 3: Performance assessment of DMHG-LEDS models on the IEMOCAP dataset.

Method	Happy	Sad	Neutral	Angry	Excited	Frustrated	ACC	W-F1
Graph-based MERC methods								
MMGCN [9]	47.10	81.91	66.44	63.51	76.17	59.06	67.10	66.81
MM-DFN [10]	43.36	83.23	70.03	70.19	73.11	64.01	69.44	68.83
M3Net [12]	60.93	78.84	70.14	68.06	77.11	67.42	70.92	71.07
GraphCFC [11]	43.08	**84.99**	64.70	**71.35**	78.86	63.70	69.13	68.91
HiMul-LGG [14]	53.95	79.92	71.66	67.56	72.00	64.00	70.12	70.22
Other representative MERC methods								
EmoLR [36]	–	–	–	–	–	–	68.53	68.12
SEDC [38]	–	–	–	–	–	–	71.60	71.68
MAGTKD [41]	–	–	–	–	–	–	69.38	69.59
ECERC [40]	60.86	79.28	71.95	66.27	78.29	67.25	71.78	71.54
IMDNet [37]	–	–	–	–	–	–	71.78	71.73
RL-EMO [39]	47.38	79.41	67.94	61.67	69.50	63.10	67.35	66.55
Ours	**61.64±0.21**	82.77±0.24	**72.48±0.19**	69.54±0.23	**81.25±0.25**	**67.82±0.22**	**73.20*±0.18**	**73.29*±0.16**

Note: MMGCN, Multimodal Fusion via Deep Graph Convolution Network for Emotion Recognition in Conversation; MM-DFN, Multimodal Dynamic Fusion Network for Emotion Recognition in Conversations; M3Net, Multivariate, Multi-frequency and Multimodal: Rethinking Graph Neural Networks for Emotion Recognition in Conversation; GraphCFC, A Directed Graph Based Cross-Modal Feature Complementation Approach for Multimodal Conversational Emotion Recognition; HiMul-LGG, A Hierarchical Decision Fusion-Based Local-Global Graph Neural Network for Multimodal Emotion Recognition in Conversation; EmoLR, Learning More from Mixed Emotions: A Label Refinement Method for Emotion Recognition in Conversations; SEDC, Semantic and Emotional Dual Channel for Emotion Recognition in Conversation; MAGTKD, Multi-modal Anchor Gated Transformer with Knowledge Distillation for Emotion Recognition in Conversation; ECERC, Evidence-Cause Attention Network for Multi-Modal Emotion Recognition in Conversation; IMDNet, Identity and Modality Attributes Driven Multimodal Fusion Networks for Emotion Recognition in Conversations; RL-EMO, Reinforcement Learning Framework for Multimodal Emotion Recognition. All values are reported in %. "–" indicates that the corresponding class-wise F1 score was not reported in the original study. Results of Joint Differentiated Modality-Aware Heterogeneous Graph and Local Emotion Difference Supervision (DMHG-LEDS) are reported as mean ± standard deviation over 10 runs. * indicates a statistically significant improvement over the reproduced MM-DFN baseline ($p < 0.05$). Boldface denotes the best result.

Table 4: Performance assessment of DMHG-LEDS models on the MELD dataset.

Method	Neutral	Surprise	Fear	Sadness	Joy	Disgust	Anger	ACC	W-F1
Graph-based MERC methods									
MMGCN [9]	78.65	57.00	17.85	**43.08**	61.11	19.19	55.14	64.67	64.95
MM-DFN [10]	78.91	57.57	21.53	42.16	62.37	25.50	54.75	65.28	65.45
M3Net [12]	79.31	57.76	20.51	40.46	63.21	25.19	52.53	66.59	65.34
DER-GCN [35]	**80.60**	51.00	10.40	41.50	64.30	10.30	**57.40**	66.80	66.10
GraphCFC [11]	78.10	56.16	21.73	37.76	60.73	24.76	53.86	64.74	64.18
HiMul-LGG [14]	–	–	–	–	–	–	–	66.21	65.18
Other representative MERC methods									
EmoLR [36]	–	–	–	–	–	–	–	66.69	65.16

(Continued)

Table 4 (continued)

Method	Neutral	Surprise	Fear	Sadness	Joy	Disgust	Anger	ACC	W-F1
SEDC [38]	–	–	–	–	–	–	–	67.43	66.16
MAGTKD [41]	–	–	–	–	–	–	–	66.36	65.32
IMDNet [37]	–	–	–	–	–	–	–	67.55	66.09
RL-EMO [39]	78.81	56.49	24.44	40.78	62.33	23.48	54.48	65.04	65.14
Ours	80.40±0.18	**59.58±0.26**	**28.87±0.24**	37.82±0.22	**65.26±0.27**	**25.53±0.23**	54.15±0.28	**68.12*±0.21**	**66.55*±0.19**

Note: MMGCN, Multimodal Fusion via Deep Graph Convolution Network for Emotion Recognition in Conversation; MM-DFN, Multimodal Dynamic Fusion Network for Emotion Recognition in Conversations; M3Net, Multivariate, Multi-frequency and Multimodal: Rethinking Graph Neural Networks for Emotion Recognition in Conversation; DER-GCN, Dialog and Event Relation-Aware Graph Convolutional Neural Network for Multimodal Dialog Emotion Recognition; GraphCFC, A Directed Graph Based Cross-Modal Feature Complementation Approach for Multimodal Conversational Emotion Recognition; HiMul-LGG, A Hierarchical Decision Fusion-Based Local-Global Graph Neural Network for Multimodal Emotion Recognition in Conversation; EmoLR, Learning More from Mixed Emotions: A Label Refinement Method for Emotion Recognition in Conversations; SEDC, Semantic and Emotional Dual Channel for Emotion Recognition in Conversation; MAGTKD, Multi-modal Anchor Gated Transformer with Knowledge Distillation for Emotion Recognition in Conversation; IMDNet, Identity and Modality Attributes Driven Multimodal Fusion Networks for Emotion Recognition in Conversations; RL-EMO, Reinforcement Learning Framework for Multimodal Emotion Recognition. All values are reported in %. "–" indicates that the corresponding class-wise F1 score was not reported in the original study. Results of DMHG-LEDS are reported as mean ± standard deviation over 10 runs. * indicates a statistically significant improvement over the reproduced MM-DFN baseline ($p < 0.05$). Boldface denotes the best result.

Table 5: Performance assessment of DMHG-LEDS models on the CMU-MOSEI dataset.

Method	HN	N	WN	Neutral	WP	P	HP	ACC	W-F1
Graph-based MERC methods									
MMGCN [9]	0.00	19.51	43.75	**42.45**	54.64	36.13	0.00	45.67	44.11
MM-DFN [10]	0.00	16.98	37.94	39.64	56.58	32.51	**8.51**	45.29	42.98
M3Net [12]	0.00	12.50	37.26	33.29	56.10	33.94	0.00	43.67	41.12
Other representative MERC methods									
DialogueCRN [43]	0.00	4.29	7.98	25.09	51.80	3.22	0.00	37.88	26.55
SACL-LSTM [42]	0.00	0.00	0.00	17.87	55.28	0.00	0.00	38.60	25.95
Ours	**16.00±0.21**	**31.76±0.19**	**45.14±0.23**	41.39±0.27	**58.14±0.18**	**37.93±0.14**	7.69**±0.22**	**47.11*±0.17**	**46.54*±0.18**

Note: MMGCN, Multimodal Fusion via Deep Graph Convolution Network for Emotion Recognition in Conversation; MM-DFN, Multimodal Dynamic Fusion Network for Emotion Recognition in Conversations; M3Net, Multivariate, Multi-frequency and Multimodal: Rethinking Graph Neural Networks for Emotion Recognition in Conversation; DialogueCRN, Contextual Reasoning Networks for Emotion Recognition in Conversations; SACL-LSTM, Supervised Adversarial Contrastive Learning for Emotion Recognition in Conversations. All values are reported in %. HN = Highly Negative, WN = Weakly Negative, WP = Weakly Positive, HP = Highly Positive. Results of DMHG-LEDS are reported as mean ± standard deviation over 10 runs. * indicates a statistically significant improvement over the reproduced MM-DFN baseline ($p < 0.05$). Boldface denotes the best result.

4.5 Ablation Study

4.5.1 Impact of Component Modules

To validate the effectiveness of each component, we conducted ablation experiments on the IEMOCAP, MELD, and CMU-MOSEI datasets. Detailed results and their visualizations are presented in Table 6. When the differentiated modality-aware graph construction strategy is removed (Group A2), the W-F1 scores decrease by 1.39%, 0.54%, and 1.76% on IEMOCAP, MELD, and CMU-MOSEI, respectively, indicating that constructing differentiated neighborhood relationships for different modalities helps enhance the effectiveness of multimodal emotion representation. After removing the entropy-gated local emotion difference supervision module (Group A3), the W-F1 scores decrease by 1.07%, 0.50%, and 1.20% on the three datasets, demonstrating that local emotion difference supervision can provide effective auxiliary discriminative information for the primary emotion classification task. When Chebyshev-KAN is replaced with a standard MLP with the same hidden dimensions, Tanh activation function, and Dropout settings (Group A4), the W-F1 drops by 0.91%, 0.15%, and 0.68% on IEMOCAP, MELD, and CMU-MOSEI, respectively. Using Transformer-FFN (Group A5) and FourierKAN (Group A6) as alternatives, the model performance surpasses the MLP replacement but still remains below that of the full model. Specifically, Transformer-FFN achieves W-F1 scores that are 0.63%, 0.12%, and 0.46% lower than the full model on the three datasets, while FourierKAN achieves scores that are 0.50%, 0.08%, and 0.35% lower. These results suggest that Chebyshev-KAN can more effectively perform nonlinear feature transformation after graph aggregation in our framework. The full model (Group A1) achieves the best results on all three datasets, indicating that differentiated graph construction, Chebyshev-KAN-based nonlinear transformation, and entropy-gated local emotion difference supervision have complementary effects, jointly improving the model's performance across different multimodal emotion recognition settings.

Table 6: Ablation results of key components and different nonlinear transformation mechanisms on IEMOCAP, MELD, and CMU-MOSEI datasets. A1 denotes the full model; A2 and A3 denote the removal of differentiated modality-aware graph construction and entropy-gated local emotion difference supervision, respectively; A4, A5, and A6 denote replacing Chebyshev-KAN with MLP, Transformer-FFN, and FourierKAN, respectively.

Group	Settings	IEMOCAP		MELD		CMU-MOSEI	
		ACC (%)	W-F1 (%)	ACC (%)	W-F1 (%)	ACC (%)	W-F1 (%)
A1	Full Model	**73.20**	**73.29**	**68.12**	**66.55**	**47.11**	**46.54**
A2	w/o Differentiated Graph	71.97	71.90	67.66	66.01	45.89	44.78
A3	w/o LEDS	72.17	72.22	67.02	66.05	46.15	45.34
A4	MLP	72.35	72.38	67.13	66.40	46.58	45.86
A5	Transformer-FFN	72.58	72.66	67.41	66.43	46.73	46.08
A6	FourierKAN	72.71	72.79	67.56	66.47	46.84	46.19

Note: ACC, Accuracy; W-F1, Weighted F1-score; w/o, without; LEDS, Entropy-Gated Local Emotion Difference Supervision; MLP, Multilayer Perceptron; FFN, Feed-Forward Network; KAN, Kolmogorov-Arnold Network. A1 denotes the full model; A2 and A3 denote the removal of differentiated modality-aware graph construction and entropy-gated local emotion difference supervision, respectively; A4, A5, and A6 denote replacing Chebyshev-KAN with MLP, Transformer-FFN, and FourierKAN, respectively. "w/o" denotes "without". Boldface denotes the best result.

4.5.2 Ablation of Modality-Specific Graph Construction

To evaluate the modality-specific graph construction, we replace one intra-modal edge set at a time with fully connected edges, while keeping the remaining intra-modal edge sets, cross-modal edges, propagation

architecture, and training settings unchanged. The results are reported in Table 7. Replacing the text position-window graph with fully connected text edges (B2) reduces the W-F1 score by 1.53% on IEMOCAP, 0.69% on MELD, and 1.18% on CMU-MOSEI. This result suggests that unrestricted aggregation over all textual utterances is less effective than preserving local discourse neighborhoods under the current setting. Replacing the visual similarity-driven graph with fully connected visual edges (B3) decreases the W-F1 score by 1.16%, 0.41%, and 1.06%, respectively, indicating that visually related nodes provide more useful contextual evidence than indiscriminate visual aggregation. Replacing the sequential audio graph with fully connected audio edges (B4) decreases the W-F1 score by 0.51%, 0.49%, and 1.11%, respectively, supporting the usefulness of short-range sequential acoustic relations. In addition, removing cross-modal edges (B5) also reduces performance on all three datasets, indicating that modality-specific intra-modal neighborhoods and cross-modal temporal alignment provide complementary contextual information. Overall, the ablation results on all three datasets support the suitability of the proposed text, visual, and audio connection rules for the present MERC setting.

Table 7: Ablation results of modality-specific graph construction on IEMOCAP, MELD, and CMU-MOSEI. In B2–B4, only the text, visual, or audio intra-modal edge set is replaced by fully connected edges, respectively, while all remaining graph components and training settings are unchanged. B5 removes cross-modal edges while retaining all intra-modal edge sets.

Group	Settings	IEMOCAP		MELD		CMU-MOSEI	
		ACC (%)	W-F1 (%)	ACC (%)	W-F1 (%)	ACC (%)	W-F1 (%)
B1	Full Model	**73.20**	**73.29**	**68.12**	**66.55**	**47.11**	**46.54**
B2	Text: fully connected	71.60	71.76	67.20	65.86	46.45	45.36
B3	Visual: fully connected	72.12	72.13	67.60	66.14	46.36	45.48
B4	Audio: fully connected	72.80	72.78	67.51	66.06	46.28	45.43
B5	Without cross-modal edges	68.52	68.26	67.36	66.07	46.13	45.26

Note: ACC, Accuracy; W-F1, Weighted F1-score; w/o, without; For B2–B4, only the indicated intra-modal edge set is replaced by fully connected edges. All values are reported in %. Boldface denotes the best result.

4.5.3 Time and Memory Overhead

Table 8 reports the practical runtime and GPU memory usage of our model and representative baselines on the MERC task. "Time" denotes the total duration of the training and testing stages in each epoch, where the testing stage corresponds to forward inference on the test set, and "Memory" denotes the allocated and reserved GPU memory usage during training. As shown in Table 8, DialogueCRN and MM-DFN exhibit longer running times across multiple datasets. Taking CMU-MOSEI as an example, their running times are 37.96 and 38.97 s, respectively. MMGCN has lower memory overhead on all three datasets, but its recognition performance falls below that of our model. M3Net and SACL-LSTM have relatively high memory consumption on some datasets. For instance, M3Net's memory usage on IEMOCAP and CMU-MOSEI reaches 12841.27 and 4549.76 MB, respectively. In contrast, DMHG-LEDS maintains relatively stable running overhead across different datasets. On IEMOCAP, our model's running time is lower than M3Net, and its memory usage is also lower than M3Net. On MELD, our model's running time is lower than DialogueCRN, MMGCN, MM-DFN, and SACL-LSTM. On CMU-MOSEI, our model's running time is lower than DialogueCRN, MMGCN, and MM-DFN, and its memory usage is lower than M3Net. Overall, although DMHG-LEDS does not have the lowest time or memory overhead on all datasets, it avoids excessive computational resource consumption while achieving favorable recognition performance, maintaining a good balance between performance and computational cost.

Table 8: Time and memory overhead comparison on IEMOCAP, MELD, and CMU-MOSEI. "Time" refers to the time required for training and testing per epoch (in seconds), and "Memory" refers to allocated and reserved memory usage (in megabytes).

Method	IEMOCAP		MELD		CMU-MOSEI	
	Time (s)	Memory (MB)	Time (s)	Memory (MB)	Time (s)	Memory (MB)
DialogueCRN	5.92	2411.95	16.58	557.15	37.96	766.55
MMGCN	1.91	547.13	14.97	191.24	30.26	344.52
MM-DFN	2.56	1939.07	23.86	767.48	38.97	680.09
SACL-LSTM	3.16	3353.32	14.50	2042.30	15.86	1634.85
M3Net	6.88	12,841.27	5.21	761.19	12.34	4549.76
DMHG-LEDS	3.45	8269.83	6.15	977.95	19.26	2288.13

Note: DialogueCRN, Contextual Reasoning Networks for Emotion Recognition in Conversations; MMGCN, Multimodal Fusion via Deep Graph Convolution Network for Emotion Recognition in Conversation; MM-DFN, Multimodal Dynamic Fusion Network for Emotion Recognition in Conversations; SACL-LSTM, Supervised Adversarial Contrastive Learning for Emotion Recognition in Conversations; M3Net, Multivariate, Multi-frequency and Multimodal: Rethinking Graph Neural Networks for Emotion Recognition in Conversation; DMHG-LEDS, Joint Differentiated Modality-Aware Heterogeneous Graph and Local Emotion Difference Supervision for Multimodal Emotion Recognition in Conversations.

4.5.4 Qualitative Analysis

To further qualitatively analyze the category discrimination characteristics of DMHG-LEDS, Fig. 3 presents the row-normalized confusion matrices of the model on IEMOCAP, MELD, and CMU-MOSEI. The diagonal regions reflect the model's recognition performance for each category, while the off-diagonal regions reveal the main confusion relationships between different emotion categories. On IEMOCAP, the predictions for Sad, Neutral, and Excited are primarily concentrated on their corresponding true categories, indicating that the model can effectively distinguish these emotions with relatively clear expressive characteristics. The main misclassifications occur between Happy and Excited, Angry and Frustrated, as well as Frustrated and Neutral, suggesting that these emotion categories share certain similarities in local semantics, intonation, or expression intensity. On MELD, the prediction distributions for Neutral and Joy are relatively concentrated, while Fear, Sadness, and Disgust are more likely to be predicted as Neutral. This phenomenon indicates that under multi-party dialogue and complex background conditions, emotions with weaker expressions or fewer samples are more easily overshadowed by the neutral state. On CMU-MOSEI, the model's confusion mainly concentrates between adjacent emotion intensity categories. For instance, some samples of Neutral, P, and HP are predicted as WP, demonstrating that the continuity of fine-grained sentiment tendencies increases the difficulty of discriminating category boundaries. Overall, Fig. 3 intuitively presents the category recognition characteristics and main confusion patterns of DMHG-LEDS under different data settings, providing qualitative explanations for the model's prediction behavior.

4.5.5 Visualization of Embedding Representations for Different Ablation Variants

To intuitively demonstrate the utterance-level emotion representations learned by different model variants, we adopt t-SNE to visualize the final fused features on the test sets of IEMOCAP, MELD, and CMU-MOSEI. As shown in Fig. 4, for each dataset, we present the representation distributions under five settings: initial, w/o Differentiated Graph, w/o Chebyshev-KAN, w/o LEDS, and Full Model. On IEMOCAP, the full model can distinguish among emotion categories more clearly, exhibiting more compact intra-class distributions and less inter-class overlap than each ablation variant. A similar phenomenon is observed

on MELD: after removing any key component, the distributions of some emotion categories become more entangled, whereas the full model obtains relatively clearer category structures. For CMU-MOSEI, some overlap still exists between adjacent emotion-intensity categories, reflecting the inherent difficulty of fine-grained sentiment orientation recognition; nevertheless, compared with each ablation variant, the full model still shows better category separation. The above visualization results indicate that the full model is capable of learning more discriminative utterance-level multimodal emotion representations.

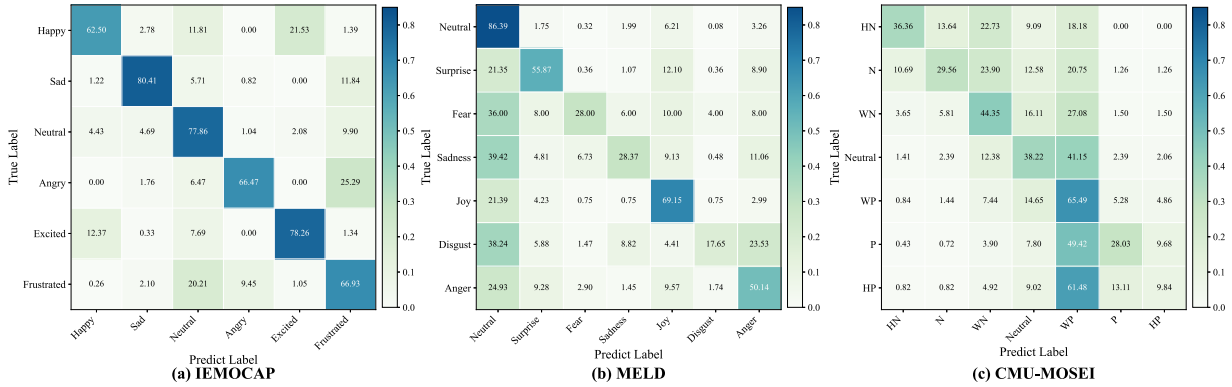


Figure 3: Row-normalized confusion matrices of Joint Differentiated Modality-Aware Heterogeneous Graph and Local Emotion Difference Supervision for Multimodal Emotion Recognition in Conversations (DMHG-LEDS) on Interactive emotional dyadic motion capture database (IEMOCAP), A Multimodal Multi-Party Dataset for Emotion Recognition in Conversation (MELD), and Multimodal Language Analysis in the Wild: CMU-MOSEI Dataset and Interpretable Dynamic Fusion Graph (CMU-MOSEI). (**a**) IEMOCAP. (**b**) MELD. (**c**) CMU-MOSEI. The horizontal axis denotes the predicted labels, and the vertical axis denotes the true labels. The color intensity indicates the normalized prediction proportion.

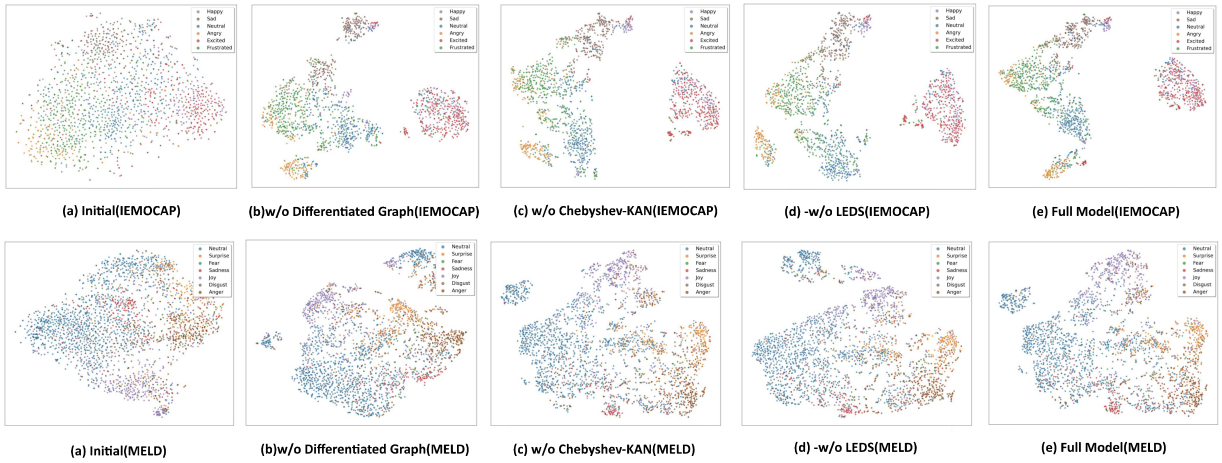


Figure 4: (Continued)

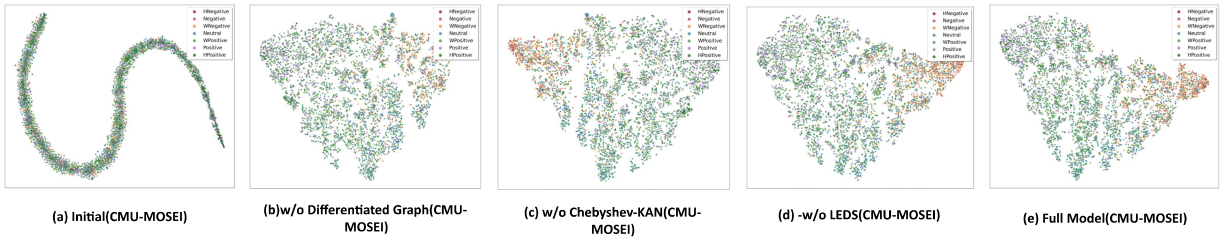


Figure 4: t-SNE visualization of utterance-level fused representations for different model variants on Interactive emotional dyadic motion capture database (IEMOCAP), A Multimodal Multi-Party Dataset for Emotion Recognition in Conversation (MELD), and Multimodal Language Analysis in the Wild: CMU-MOSEI Dataset and Interpretable Dynamic Fusion Graph (CMU-MOSEI). The first, second, and third rows correspond to IEMOCAP, MELD, and CMU-MOSEI, respectively; from left to right: (**a**) initial; (**b**) w/o Differentiated graph; (**c**) w/o Chebyshev Polynomial-Based Kolmogorov-Arnold Networks: An Efficient Architecture for Nonlinear Function Approximation (Chebyshev-KAN); (**d**) w/o LEDS; (**e**) Full model.

5 Conclusion

This paper proposes DMHG-LEDS for multimodal emotion recognition in conversations. The framework integrates differentiated modality-aware graph construction, Chebyshev-KAN-based graph propagation, and entropy-gated local emotion difference supervision to enhance utterance representation learning. Experiments on IEMOCAP, MELD, and CMU-MOSEI demonstrate the effectiveness of the proposed method.

Although DMHG-LEDS has been validated on standard benchmarks with complete multimodal inputs, extending the framework to more complex application scenarios remains an important direction for future research. Noise in visual or acoustic signals may increase the uncertainty of visual similarity estimation and cross-modal feature aggregation. Under missing-modality conditions, how to maintain stable emotional representations while preserving effective cross-modal interactions also warrants further investigation. Moreover, as conversation length increases, both the graph scale and the number of candidate similarity relations grow accordingly. The current text and cross-modal windows can be further optimized to better model long-range emotional dependencies. Future work will explore reliability-aware relation modeling, robust representation learning for incomplete modalities, and efficient long-range context modeling.

Acknowledgement: Not applicable.

Funding Statement: This work is supported by the National Natural Science Foundation of China under Grants 61602161 and 61772180, Hubei Province Science and Technology Support Project under Grant 2020BAB012, and the Fundamental Research Funds for the Research Fund of Hubei University of Technology under Grants HBUT: 2021046, 21060, and 21066.

Author Contributions: The authors confirm contribution to the paper as follows: study conception and design: Yu Chen, Panpan Chen and Qun Zhang; data collection and experimental implementation: Yu Chen, Jiahui Huang and Xinyi Zhu; analysis and interpretation of results: Yu Chen, Panpan Chen, Jun Wu and Shuai Guo; draft manuscript preparation: Yu Chen; manuscript review and editing: Panpan Chen, Jun Wu, Shuai Guo, Jiahui Huang, Xinyi Zhu and Qun Zhang; supervision: Jun Wu and Qun Zhang. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: The datasets used during the current study are available. Additionally, the datasets generated, model settings, and training processes are available from the corresponding author upon reasonable request.

Ethics Approval: The authors state that this research complies with ethical standards. This research does not involve either human participants or animals.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Zhang X, Cui W, Hu B, Li Y. A multi-level alignment and cross-modal unified semantic graph refinement network for conversational emotion recognition. IEEE Trans Affect Comput. 2024;15(3):1553–66. doi:10.1109/TAFFC.2024.3354382.
2. Shou Y, Meng T, Ai W, Zhang F, Yin N, Li K. Adversarial alignment and graph fusion via information bottleneck for multimodal emotion recognition in conversations. Inf Fusion. 2024;112:102590. doi:10.1016/j.inffus.2024.102590.
3. Wang C, Liao M, Huang Z, Wu J, Zong C, Zhang J. BLSP-Emo: towards empathetic large speech-language models. In: Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing; 2024 Nov 12–16; Miami, FL, USA. p. 19186–99. doi:10.18653/v1/2024.emnlp-main.1070.
4. Zou HP, Huang WC, Wu Y, Chen Y, Miao C, Nguyen H, et al. LLM-based human-agent collaboration and interaction systems: a survey. arXiv:2505.00753. 2025. doi:10.48550/arXiv.2505.00753.
5. Meng T, Zhang F, Shou Y, Ai W, Yin N, Li K. Revisiting multimodal emotion recognition in conversation from the perspective of graph spectrum. arXiv:2404.17862. 2024. doi:10.48550/arXiv.2404.17862.
6. Gan C, Zheng J, Zhu Q, Cao Y, Zhu Y. A survey of dialogic emotion analysis: developments, approaches and perspectives. Pattern Recognit. 2024;156(3):110794. doi:10.1016/j.patcog.2024.110794.
7. Jun W, Tianliang Z, Jiahui Z, Tianyi L, Chunzhi W. Hierarchical multiples self-attention mechanism for multi-modal analysis. Multimed Syst. 2023;29(6):3599–608. doi:10.1007/s00530-023-01133-7.
8. Wu J, Wang J, Jing S, Liu J, Zhang T, Han M, et al. Text-dominant strategy for multistage optimized modality fusion in multimodal sentiment analysis. Multimed Syst. 2024;30(6):353. doi:10.1007/s00530-024-01518-2.
9. Hu J, Liu Y, Zhao J, Jin Q. MMGCN: multimodal fusion via deep graph convolution network for emotion recognition in conversation. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing; 2021 Aug 1–6; Virtual. p. 5666–75. doi:10.18653/V1/2021.ACL-LONG.440.
10. Hu D, Hou X, Wei L, Jiang L, Mo Y. MM-DFN: multimodal dynamic fusion network for emotion recognition in conversations. In: Proceedings of the ICASSP 2022—2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); 2022 May 23–27; Singapore. p. 7037–41. doi:10.1109/ICASSP43922.2022.9747397.
11. Li J, Wang X, Lv G, Zeng Z. GraphCFC: a directed graph based cross-modal feature complementation approach for multimodal conversational emotion recognition. IEEE Trans Multimed. 2024;26:77–89. doi:10.1109/TMM.2023.3260635.
12. Chen F, Shao J, Zhu S, Shen HT. Multivariate, multi-frequency and multimodal: rethinking graph neural networks for emotion recognition in conversation. In: Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2023 Jun 17–24; Vancouver, BC, Canada. p. 10761–70. doi:10.1109/CVPR52729.2023.01036.
13. Li J, Wang X, Zeng Z. Tracing intricate cues in dialogue: joint graph structure and sentiment dynamics for multimodal emotion recognition. IEEE Trans Pattern Anal Mach Intell. 2025;47(10):8786–803. doi:10.1109/TPAMI.2025.3581236.
14. Fu C, Qian F, Su K, Su Y, Wang Z, Shi J, et al. HiMul-LGG: a hierarchical decision fusion-based local-global graph neural network for multimodal emotion recognition in conversation. Neural Netw. 2025;181(4):106764. doi:10.1016/j.neunet.2024.106764.
15. Meng T, Shou Y, Ai W, Du J, Liu H, Li K. A multi-message passing framework based on heterogeneous graphs in conversational emotion recognition. Neurocomputing. 2024;569(1):127109. doi:10.1016/j.neucom.2023.127109.
16. Ghosal D, Majumder N, Poria S, Chhaya N, Gelbukh A. DialogueGCN: a graph convolutional neural network for emotion recognition in conversation. In: Proceedings of the 2019 Conference on Empirical Methods in Natural

Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP); 2019 Nov 3–7; Hong Kong, China. p. 154–64. doi:10.18653/V1/D19-1015.
17. Gao Q, Cao B, Guan X, Gu T, Bao X, Wu J, et al. Emotion recognition in conversations with emotion shift detection based on multi-task learning. Knowl Based Syst. 2022;248(8):108861. doi:10.1016/j.knosys.2022.108861.
18. Agarwal H, Bansal K, Joshi A, Modi A. Shapes of emotions: multimodal emotion recognition in conversations via emotion shifts. arXiv:2112.01938. 2022.
19. Tu G, Xiong F, Liang B, Xu R. A persona-infused cross-task graph network for multimodal emotion recognition with emotion shift detection in conversations. In: Proceedings of the SIGIR'24: Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval; 2024 Jul 14–18; Washington, DC, USA. p. 2266–70. doi:10.1145/3626772.3657944.
20. Sidharth SS, Keerthana AR, Gokul R, Anas KP. Chebyshev polynomial-based Kolmogorov-Arnold Networks: an efficient architecture for nonlinear function approximation. arXiv:2405.07200. 2024.
21. Wang R, Guo C, Shabaz M, Rida I, Cambria E, Zhu X. CIME: contextual interaction-based multimodal emotion analysis with enhanced semantic information. IEEE Trans Comput Soc Syst. 2026;13(3):4001–11. doi:10.1109/TCSS.2025.3572495.
22. Tellamekala MK, Amiriparian S, Schuller BW, André E, Giesbrecht T, Valstar M. COLD fusion: calibrated and ordinal latent distribution fusion for uncertainty-aware multimodal emotion recognition. IEEE Trans Pattern Anal Mach Intell. 2024;46(2):805–22. doi:10.1109/TPAMI.2023.3325770.
23. Sun L, Lian Z, Liu B, Tao J. Efficient multimodal transformer with dual-level feature restoration for robust multimodal sentiment analysis. IEEE Trans Affect Comput. 2024;15(1):309–25. doi:10.1109/TAFFC.2023.3274829.
24. Guo Z, Jin T, Zhao Z. Multimodal prompt learning with missing modalities for sentiment analysis and emotion recognition. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics; 2024 Aug 11–16; Bangkok, Thailand. p. 1726–36. doi:10.18653/v1/2024.acl-long.94.
25. Zhu X, Feng H, Cambria E, Yu X, Santamaria J, Fan X, et al. FMSA: few-shot multimodal sentiment analysis for social media via integrated prompt learning and vision-language models. IEEE Trans Affect Comput. 2026. doi:10.1109/TAFFC.2026.3681216.
26. Simić N, Suzić S, Milošević N, Stanojev V, Nosek T, Popović B, et al. Enhancing emotion recognition through federated learning: a multimodal approach with Convolutional Neural Networks. Appl Sci. 2024;14(4):1325. doi:10.3390/app14041325.
27. Shen W, Wu S, Yang Y, Quan X. Directed acyclic graph network for conversational emotion recognition. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing; 2021 Aug 1–6; Bangkok, Thailand. p. 1551–60. doi:10.18653/v1/2021.acl-long.123.
28. Liu Z, Wang Y, Vaidya S, Ruehle F, Halverson J, Soljačić M, et al. KAN: Kolmogorov-Arnold Networks. arXiv:2404.19756. 2024. doi: 10.48550/arxiv.2404.19756.
29. Bozorgasl Z, Chen H. Wav-KAN: Wavelet Kolmogorov-Arnold Networks. arXiv:2405.12832. 2024. doi:10.48550/arXiv.2405.12832.
30. Mehrabian A, Adi PM, Heidari M, Hacihaliloglu I. Implicit neural representations with fourier kolmogorov-arnold networks. arXiv:2409.09323. 2024. doi: 10.48550/arxiv.2409.09323.
31. Li J, Wang X, Liu Y, Zeng Z. CFN-ESA: a cross-modal fusion network with emotion-shift awareness for dialogue emotion recognition. IEEE Trans Affect Comput. 2024;15(4):1919–33. doi:10.1109/TAFFC.2024.3389453.
32. Busso C, Bulut M, Lee CC, Kazemzadeh A, Mower E, Kim S, et al. IEMOCAP: interactive emotional dyadic motion capture database. Lang Resour Eval. 2008;42(4):335–59. doi:10.1007/S10579-008-9076-6.
33. Poria S, Hazarika D, Majumder N, et al. MELD: a multimodal multi-party dataset for emotion recognition in conversations. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics; 2019 Jul 28–Aug 2; Florence, Italiy. p. 527–36. doi:10.18653/V1/P19-1050.
34. Bagher Zadeh A, Liang PP, Poria S, Cambria E, Morency LP. 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; 2018 Jul 15–20; Melbourne, Australia. p. 2236–46. doi:10.18653/v1/P18-1208.

35. Ai W, Shou Y, Meng T, Li K. DER-GCN: dialog and event relation-aware graph convolutional neural network for multimodal dialog emotion recognition. IEEE Trans Neural Netw Learn Syst. 2025;36(3):4908–21. doi:10.1109/TNNLS.2024.3367940.
36. Wen J, Tu G, Li R, Jiang D, Zhu W. Learning more from mixed emotions: a label refinement method for emotion recognition in conversations. TACL. 2023;11(4):1485–99. doi:10.1162/tacl_a_00614.
37. Shi W, Chen X, Yao B, Wen Y, Sheng B. Identity and modality attributes driven multimodal fusion networks for emotion recognition in conversations. IEEE Trans Multimed. 2025;27:4361–71. doi:10.1109/TMM.2025.3535347.
38. Yang Z, Zhang Z, Cheng Y, Wang X. Semantic and emotional dual channel for emotion recognition in conversation. IEEE Trans Affect Comput. 2025;16(3):1885–902. doi:10.1109/TAFFC.2025.3544608.
39. Zhang C, Zhang Y, Cheng B. RL-EMO: reinforcement learning framework for multimodal emotion recognition. In: Proceedings of the ICASSP 2024—2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); 2024 Apr 14–19; Seoul, Republic of Korea. p. 10246–50. doi:10.1109/ICASSP48485.2024.10446459.
40. Zhang T, Tan Z. ECERC: evidence-cause attention network for multi-modal emotion recognition in conversation. In: Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics; 2025 Jul 27; Vienna, Austria. p. 2064–77. doi:10.18653/v1/2025.acl-long.102.
41. Li J, Ding S, Guo L, Li X. Multi-modal anchor gated transformer with knowledge distillation for emotion recognition in conversation. In: Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence; 2025 Aug 16–22; Montreal, Canada. p. 8141–9. doi:10.24963/ijcai.2025/905.
42. Hu D, Bao Y, Wei L, Zhou W, Hu S. Supervised adversarial contrastive learning for emotion recognition in conversations. In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics; 2023 Jul 9–14; Toronto, Canada. p. 10835–52. doi:10.18653/v1/2023.acl-long.606.
43. Hu D, Wei L, Huai X. DialogueCRN: contextual reasoning networks for emotion recognition in conversations. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing; 2021 Aug 1–6; Virtual. p. 7042–52. doi:10.18653/v1/2021.acl-long.547.