Expert Systems With Applications 319 (2026) 132002

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

AMB-DSGDN: Adaptive modality-balanced dynamic semantic graph differential network for multimodal emotion recognition ⋆

Yunsheng Wang [a], Yuntao Shou [a], Yilong Tan [a], Wei Ai [a], Tao Meng [a,*], Keqin Li [b]

[a] College of Computer and Mathematics, Central South University of Forestry and Technology, 410004, Hunan Changsha, China
[b] Department of Computer Science, State University of New York, New Paltz, New York, 12561, USA

ARTICLE INFO

Keywords:
Multimodal emotion recognition
Adaptive modality balancing
Speaker semantic graph
Graph neural networks

ABSTRACT

Multimodal dialogue emotion recognition captures emotional cues by fusing text, visual, and audio modalities. However, existing approaches still suffer from notable limitations in modeling emotional dependencies and learning multimodal representations. On the one hand, they are unable to effectively filter out redundant or noisy signals within multimodal features, which hinders the accurate capture of the dynamic evolution of emotional states across and within speakers. On the other hand, during multimodal feature learning, dominant modalities (e.g., textual cues) tend to overwhelm the fusion process, thereby suppressing the complementary contributions of non-dominant modalities such as speech and vision, ultimately constraining the overall recognition performance. To address these challenges, we propose an Adaptive Modality-Balanced Dynamic Semantic Graph Differential Network (AMB-DSGDN). Concretely, we first construct modality-specific subgraphs for text, speech, and vision, where each modality contains intra-speaker and inter-speaker graphs to capture both self-continuity and cross-speaker emotional dependencies. On top of these subgraphs, we introduce a differential graph attention mechanism, which computes the discrepancy between two sets of attention maps. By explicitly contrasting these attention distributions, the mechanism cancels out shared noise patterns while retaining modality-specific and context-relevant signals, thereby yielding purer and more discriminative emotional representations. In addition, we design an adaptive modality balancing mechanism, which estimates a dropout probability for each modality according to its relative contribution in emotion modeling. This mechanism randomly discards a portion of features from dominant modalities to suppress their overwhelming influence, while proportionally rescaling the preserved features based on the dropout probability to maintain overall information balance. Extensive experiments on IEMOCAP and MELD datasets validate that AMB-DSGDN significantly outperforms state-of-the-art baselines, demonstrating its effectiveness and robustness in multimodal conversational emotion recognition.

1. Introduction

Dialogue emotion recognition is a key task in human-computer interaction, natural language processing, and affective computing, aiming to accurately identify speakers' emotional states in multi-party dialogues to enhance the performance of intelligent systems in applications such as social robots, virtual assistants, mental health monitoring, and customer service (Ai et al., 2025; Poria et al., 2018; Wu et al., 2025a). With the rapid development of artificial intelligence, dialogue emotion recognition has evolved from unimodal to multimodal approaches, integrating text, visual, and audio sources to capture the complexity and multidimensional features of human emotional expression (Tu et al., 2024; Wu et al., 2025b). For instance, in everyday conversations, emotions are conveyed not only through verbal content but also through facial expressions and vocal tones, where the complementarity of these modalities helps models better understand emotional dynamics (Sun & Zhou, 2025).

It is worth further attention that emotional states in dialogues evolve continuously with the interaction process, exhibiting significant dynamic characteristics (Fu et al., 2025). In real scenarios, individuals may sustain prior emotional states or adjust instantly due to others' influences, reflecting intra-speaker continuity and inter-speaker interactivity (Mehrez & Selouani, 2025). Meanwhile, the importance of different modalities fluctuates dynamically during the dialogue, affecting the model's ability to capture emotional changes (Lian et al., 2023). Fig. 1 illustrates a typical dialogue segment. The male, as a staff member,

⋆ Our code is publicly available at https://github.com/wys-ljq/AMB-DSGDN.

* Corresponding author.
E-mail addresses: 20234261@csuft.edu.cn (Y. Wang), shouyuntao@stu.xjtu.edu.cn (Y. Shou), tanyilong@csuft.edu.cn (Y. Tan), aiwei@hnu.edu.cn (W. Ai), mengtao@hnu.edu.cn (T. Meng), lik@newpaltz.edu (K. Li).

https://doi.org/10.1016/j.eswa.2026.132002
Received 20 September 2025; Received in revised form 11 January 2026; Accepted 6 March 2026
Available online 14 March 2026

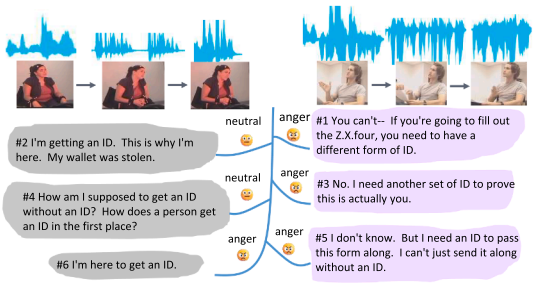

Fig. 1. An authentic and representative segment illustrating the dynamic evolution of dialogue from the IEMOCAP dataset (Ses01F_impro01).

exhibits a stern attitude from the beginning, with his anger primarily manifested through tone and facial expressions, persisting in subsequent exchanges, reflecting the continuity of intra-speaker dependency. The female initially makes a request in a natural tone, with the text modality providing rational semantic information, but as the male's sternness continues, she shifts to anger, with changes in vocal intensity and facial expressions becoming more prominent, demonstrating sensitivity to the other's emotions. This process indicates that the female's emotional changes rely more on inter-speaker emotional influences, highlighting the dynamic contagion of cross-speaker dependencies. At the same time, this segment reveals the phased dominance of different modalities in the dialogue process: the text modality typically provides clearer semantic information in the early stages, while audio and visual modalities play a larger role when emotional changes are evident.

Unlike traditional static feature modeling, multimodal dialogue emotion recognition tasks require capturing the dynamic evolution of emotion dependencies over time (Cheng et al., 2024; Zhao et al., 2025a,b). If only static graph structures are employed for modeling, emotion dependencies are often reduced to fixed relational patterns, making it difficult to capture the dynamic changes in emotions driven by contextual variations, thus limiting the ability to model complex dialogue scenarios (Farhadipour et al., 2025). Meanwhile, multimodal features commonly contain redundant and shared noise, which, if indiscriminately incorporated into the modeling process, can obscure effective emotional signals and reduce the discriminative power of representations (Fan et al., 2024). Additionally, this task must address the fluctuating contributions of different modalities throughout the dialogue process. If the model overlooks the dynamic differences between modalities, it is likely to result in dominant modalities overly influencing the outcome, while weaker modalities fail to contribute effectively, ultimately degrading overall recognition performance (Shou et al., 2024). Therefore, to effectively capture the continuously evolving emotional states in dialogues, it is essential to achieve dynamic modeling of emotion dependencies, effective suppression of noise in modeled features, and adaptive regulation of modality contributions.

To this end, researchers in multimodal dialogue emotion recognition (MERC) have proposed various modeling approaches, mainly divided into two directions: sequence structure methods rely on recurrent or memory networks to depict temporal dependencies, capable of capturing local continuity but struggling to cover long-distance inter-speaker interactions (Majumder et al., 2019); graph structure methods build dialogue graphs to model intra- and inter-speaker dependencies but mostly use static edge weights, ignoring the dynamic changes of dependency relationships with time and context (Ghosal et al., 2019). At the same time, for the fluctuating contributions of different modalities in the dialogue process, existing methods lack effective balancing regulation mechanisms, easily causing dominant modalities to prevail and weaker modalities to be diminished (Guo et al., 2024a).

To address the aforementioned limitations, this paper proposes a Dynamic Semantic Graph Differential Network with Adaptive Modal Balancing. Specifically, for modeling emotion dependencies, we construct modality-specific subgraphs for text, speech, and visual modalities to capture both inter-speaker interactions and intra-speaker emotional evolution. On top of these subgraphs, we design a graph differential attention mechanism. This mechanism first projects utterance features into a unified representation space and computes attention distributions between nodes using left and right linear transformations, while explicitly modeling inter-speaker interactions by incorporating relational embeddings. Subsequently, a differential operation is performed on the two attention distributions, canceling out their overlapping components and retaining only the differential parts. This approach effectively removes shared noise present across different modalities while emphasizing modality-specific and contextually relevant dependency signals. Through this differential modeling strategy, the model not only filters out redundant information but also captures the dynamic dependencies evolving with context in dialogues, resulting in purer and more discriminative emotional representations. For modality regulation, we introduce an adaptive modal dropout mechanism. The model calculates dropout probabilities based on the relative performance of each modality in emotion recognition and randomly discards a portion of the dominant modality's features through probabilistic sampling, while scaling the retained features to ensure the stability of the overall information content. This strategy effectively mitigates modality imbalance, preventing any single modality from overly dominating the fusion process. Through the synergistic interaction of these two mechanisms, the model can dynamically capture emotional evolution in dialogues. The main contributions of this paper are as follows:

- We propose AMB-DSGDN, which explicitly constructs modality-specific subgraphs to model both intra-speaker and inter-speaker emotional dependencies. By jointly integrating differential graph attention and an adaptive modality balancing mechanism, the model effectively captures dynamic emotional variations in conversations while alleviating noise interference and modality imbalance, enhancing the discriminability of emotional representations.
- We design a differential graph attention mechanism that computes discrepancies between paired attention maps on modality-specific subgraphs. Through differential contrast, this mechanism suppresses shared noise and highlights modality-specific and context-relevant information, improving dynamic emotion modeling capability.
- We further introduce an adaptive dropout-based modality balancing mechanism, which dynamically identifies the dominant modality and randomly drops part of its features, while proportionally rescaling the retained features, thereby alleviating the impact of single-modality dominance and enabling balanced multimodal information fusion.
- Extensive experiments on the IEMOCAP and MELD datasets demonstrate the superior performance and robustness of AMB-DSGDN.

2. Related works

This section reviews recent advances in multimodal learning, dynamic emotional dependencies, and modal imbalance learning, highlighting key challenges and research directions for improving multimodal conversational emotion recognition.

2.1. Multimodal learning

In recent years, multimodal learning has made significant progress in fields such as computer vision, natural language processing, and speech analysis. Researchers have proposed various methods to enhance multi-source information fusion and improve model generalization performance. In multimodal emotion recognition tasks, the NORM-TR model (Liu et al., 2024) effectively captures long-range dependencies between modalities and improves the model's robustness to noise and computational efficiency through a noise-robust feature extractor and noise-aware learning scheme combined with a Transformer fusion mechanism. The AffectGPT model (Lian et al., 2025) uses

a pre-fusion operation and multimodal large language model architecture, placing cross-modal interactions outside the LLM to capture fine-grained emotions from text, audio, and visual inputs. The HKD-MER model (Sun et al., 2024) enhances feature balance and discriminative ability by transferring dominant modality knowledge to other modalities through hierarchical knowledge distillation. The MERBench model (Lian et al., 2024b) achieves fine-grained modeling of modality heterogeneity and cross-modal interactions using an attention-based fusion framework. Overall, these methods have achieved positive results in fusion effects and generalization capabilities. However, existing research often fails to fully consider the dynamic differences in emotional expressions across modalities in dialogue scenarios, cannot effectively filter redundant noise in multimodal features, and the dominant modality (such as text) overly dominates the fusion process, suppressing the complementary role of non-dominant modalities, thereby limiting overall recognition performance. To this end, this paper constructs modality-specific subgraphs for dynamic emotion modeling, introduces differential graph attention to offset shared noise, and adopts an adaptive dropout strategy to adjust modality contributions, thereby improving joint learning in multimodal models.

2.2. Dynamic emotional dependencies

In the field of emotion recognition, recent research has proposed various methods to better capture dynamic emotional dependencies, mainly including two categories: recurrent neural network (RNN)-based and graph convolutional network (GCN)-based models. RNN-based methods (such as the optimized RNN model (Reddy et al., 2025)) can handle sequential data and model temporal dependencies, but are prone to gradient vanishing or explosion in long sequences, limiting the expression of complex emotional dynamics. To enhance modeling capabilities, researchers use GCN to handle emotional association structures, suitable for dialogue scenarios with multi-turn interactions. DER-GCN (Ai et al., 2024) strengthens inter-speaker dependency modeling by fusing dialogue-aware and event-aware information; SERC-GCN (Chandola et al., 2024) captures changes in speaker emotional states to better capture emotional evolution. Some studies also attempt to fuse RNN and GCN, such as GCN-LSTM (Kong et al., 2024), to balance temporal and structural information. Additionally, DEDNet (Wang et al., 2024) models inter- and intra-speaker emotional dependencies and uses interaction to capture emotion changes. Although these methods have made progress, most graph-based methods still rely on static edge weights, and noise in multimodal features weakens the characterization of dynamic emotional dependencies. Therefore, this paper constructs intra-speaker and inter-speaker subgraphs in each modality to express temporal and interactive relationships of emotions. At the same time, positive and negative branch differential attention is introduced in the graph, using the difference in their attention and amplifying stable and consistent emotional features to achieve more reliable dynamic emotion modeling.

2.3. Modal imbalance learning

Modality imbalance is one of the core challenges in multimodal learning in recent years. Different modalities differ in data quality, information density, and availability, leading to some modalities dominating the fusion while others' contributions are weakened. To address this issue, researchers have proposed various strategies. Wei et al. (2024) designed a dynamic modality assignment framework that adaptively adjusts weights for each modality, enhancing the contribution of secondary modalities and alleviating single-modality dominance. Wang et al. (2025) proposed an adversarial modality balancing method that enhances underrepresented modality features through quantity-quality reweighting, improving fusion effects. The MPLMM model proposed by Guo et al. (2024b) dynamically suppresses noisy modalities in emotion recognition while retaining key information. The MER framework proposed by Lian et al. (2024a) combines modality robustness with semi-supervised learning to enhance the discriminative ability of secondary modalities. Chen et al. (2024) use multimodal knowledge distillation in a teacher-student architecture to achieve cross-modal knowledge transfer, alleviating performance degradation caused by modality imbalance. Nevertheless, existing methods do not fully consider the fluctuations in modality contributions with contextual changes in dynamic dialogues, leading to greater model fluctuations under extreme imbalances. To this end, this paper proposes an adaptive modality balancing strategy that dynamically adjusts feature sampling by evaluating each modality's performance in the current batch and proportionally amplifies retained features, thereby moderately suppressing features of dominant modalities when they are too strong while enhancing secondary modality contributions to achieve balanced fusion of multimodal features.

3. Task definition

Multimodal Emotion Recognition in Conversation aims to identify the emotion category of each utterance in a dialogue sequence. The input to this task is a dialogue sequence comprising multiple utterances, where each utterance includes features from text, visual, and audio modalities, along with speaker information and contextual relationships. Formally, given a dialogue sequence $\mathcal{D} = \{u_1, u_2, \ldots, u_N\}$, where N is the number of utterances, each utterance u_i consists of its text feature $\mathbf{t}_i \in \mathbb{R}^{d_t}$, visual feature $\mathbf{v}_i \in \mathbb{R}^{d_v}$, audio feature $\mathbf{a}_i \in \mathbb{R}^{d_a}$, speaker identifier s_i, and emotion label $y_i \in \{1, 2, \ldots, C\}$ (C is the number of emotion categories). The model's objective is to learn a mapping function $f : \mathcal{D} \rightarrow \{y_1, y_2, \ldots, y_N\}$ to maximize prediction accuracy and weighted F1 score. This task incorporates multimodal fusion, speaker dependencies, and contextual relationships, making it suitable for datasets such as IEMOCAP and MELD.

4. Methodology

In this section, we introduce the proposed multimodal emotion recognition method for conversations. The method encompasses six key components: model architecture, utterance-level encoder, relational subgraph construction, differential attention graph convolutional network, dynamic modality balancing mechanism, and emotion classifier.

4.1. Model architecture

AMB-DSGDN combines differential attention graph convolutional networks with dynamic modality balancing mechanisms. Fig. 2 shows the overall framework of the model, mainly consisting of five core components: utterance-level encoder, relational subgraph construction, differential attention graph convolutional network, dynamic modality balancing mechanism, and emotion classifier. First, text, audio, and visual features are mapped to unified dimensional hidden representations through respective linear transformations. Subsequently, a Transformer encoder incorporating speaker embeddings is introduced for contextual modeling to capture temporal dependencies between utterances. On this basis, the model constructs intra-speaker and inter-speaker relational subgraphs and uses differential attention graph convolution mechanisms to model emotional dependencies. This mechanism suppresses shared redundancies and noise patterns by modeling differences in node attention distributions in the same modality subgraph, thereby highlighting true emotional dependency signals. This design helps more accurately characterize dynamic emotional changes in dialogues. In terms of modality regulation, the adaptive modality dropout mechanism generates dropout probabilities based on each modality's contribution to emotion modeling and scales compensated retained features to achieve dynamic weight adjustment across modalities. Finally, the fused multimodal features are fed into the emotion classifier for prediction, while auxiliary losses are introduced to enhance the robustness of unimodal emotional representations.

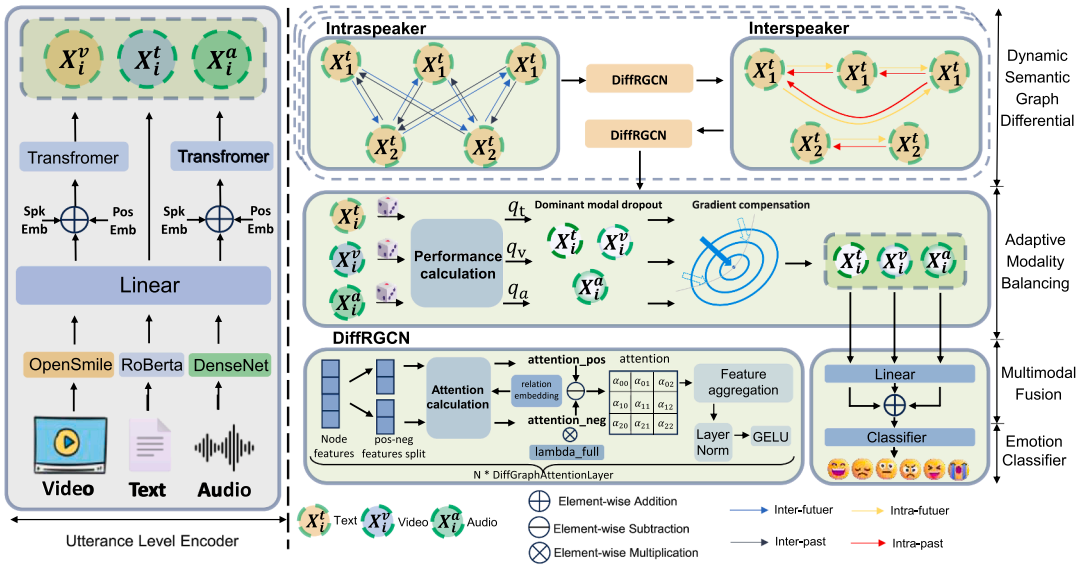

Fig. 2. This architecture includes four core modules: first, the utterance-level encoder extracts unimodal features through OpenSmile (audio), RoBERTa (text), DenseNet (video), and after Transformer combining speaker embedding (Spk Emb) and position embedding (Pos Emb) encoding, obtains the text feature $\mathbf{x}_i^t$, video feature $\mathbf{x}_i^t$, audio feature $\mathbf{x}_i^t$ for the i-th utterance; second, the differential graph attention module constructs subgraphs including "intra-speaker subgraph (Intraspeaker)" and "inter-speaker subgraph (Interspeaker)" for each modality, computes differences between two groups of attention distributions through differential graph attention, eliminates cross-modal shared noise and retains modality-specific emotional signals; then, the adaptive modality balancing module computes dropout probabilities ($q_t/q_v/q_a$) for each modality based on batch-level performance, performs dynamic dropout on dominant modalities, while scaling retained features through gradient compensation to maintain information balance; finally, through multimodal fusion and classification module, fuses the balanced features via linear layers and inputs into the classifier to obtain emotion recognition results.

4.2. Utterance-level encoder

The utterance-level encoder is used to perform feature extraction and fusion on multimodal inputs (text, visual, and audio) to generate semantic representations for each utterance in the dialogue. This module consists of modality feature projection, position encoding and speaker embedding, and Transformer encoding layers, ultimately outputting unified representations for the three modalities.

For each utterance in the dialogue, initial features are first extracted from pre-trained models: text modality uses RoBERTa (Liu et al., 2019), visual modality uses DenseNet (Huang et al., 2017), audio modality uses OpenSmile (Eyben et al., 2010). Subsequently, linear mappings project different modality features to unified hidden dimensions:

$$\mathbf{x}_i^t = W_t \cdot \mathbf{f}_i^t + \mathbf{b}_t, \quad \mathbf{x}_i^v = W_v \cdot \mathbf{f}_i^v + \mathbf{b}_v, \quad \mathbf{x}_i^a = W_a \cdot \mathbf{f}_i^a + \mathbf{b}_a, \tag{1}$$

where $\mathbf{f}_i^t, \mathbf{f}_i^v, \mathbf{f}_i^a$ represent the original text, visual, and audio features of the i-th utterance, respectively, $W_t \in \mathbb{R}^{d_h \times D_t}, W_v \in \mathbb{R}^{d_h \times D_v}, W_a \in \mathbb{R}^{d_h \times D_a}$ are projection matrices, $\mathbf{b}_t, \mathbf{b}_v, \mathbf{b}_a$ are bias terms.

To introduce sequence position information and speaker identity information, the model adds position encoding and speaker embedding. Position encoding uses sine-cosine form:

$$\begin{aligned} PE(pos, 2j) &= \sin\left(\frac{pos}{10000^{2j/d_h}}\right), \\ PE(pos, 2j+1) &= \cos\left(\frac{pos}{10000^{2j/d_h}}\right) \end{aligned} \tag{2}$$

where pos represents the position of the utterance in the sequence, j is the dimension index. Speaker embedding is generated through an embedding layer: $\mathbf{s}_i = E_s(\text{spk}_i)$,where $E_s \in \mathbb{R}^{(n_s+1)\times d_h}$,$n_s$ is the number of speakers.

For audio and visual modalities, add the projected features with position encoding and speaker embedding:

$$\mathbf{x}_i^a = \mathbf{x}_i^a + PE(i) + \mathbf{s}_i, \quad \mathbf{x}_i^v = \mathbf{x}_i^v + PE(i) + \mathbf{s}_i. \tag{3}$$

Subsequently, audio and visual modalities are input into independent Transformer encoders to model contextual dependencies. Each encoder consists of a single-layer multi-head self-attention and position-wise feed-forward network. The self-attention mechanism is defined as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V, \tag{4}$$

where $Q = W_q\mathbf{x}, K = W_k\mathbf{x}, V = W_v\mathbf{x}, d_k = d_h/h, h$ is the number of attention heads. The position-wise feed-forward network is defined as:

$$FFN(\mathbf{x}) = W_2 \cdot \text{GELU}(W_1 \cdot LN(\mathbf{x})) + \mathbf{x}, \tag{5}$$

where LN represents layer normalization.

Finally, the contextual representations of audio and visual modalities are uniformly represented as:

$$\mathbf{x}_i^m, \mathbf{A}_{attn}^m = \text{TransformerEncoder}(\mathbf{x}^m, \mathbf{mask}, \mathbf{s}), \quad m \in \{a, v\}, \tag{6}$$

where $\mathbf{A}_{attn}^m$ is the attention matrix reflecting dependencies between utterances, $\mathbf{mask}$ is the utterance mask, $\mathbf{s}$ represents sequence information.

4.3. Relational subgraph construction

In dialogues, each utterance u_i is represented as graph node v_i, with multimodal features $\mathbf{x}_i^m, m \in \{t, v, a\}$, corresponding to text, visual, and audio modalities respectively.

4.4. Graph construction and representation

A complete dialogue is represented as $U = \{u_1, u_2, \ldots, u_{N_b}\}$, and modeled as a directed graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{R}, \mathcal{W})$, where each utterance u_i corresponds to node $v_i \in \mathcal{V}$. Node features consist of multimodal representations $(\mathbf{x}_i^t, \mathbf{x}_i^v, \mathbf{x}_i^a)$, corresponding to text, visual, and audio modalities. To model emotional dependencies, the model constructs two relational subgraphs: intra-speaker subgraph $\mathbf{Adj}_s$ and inter-speaker subgraph $\mathbf{Adj}_c$, characterizing temporal evolution within the same speaker and interactive relationships between different speakers respectively. Edge set $\mathcal{E}$ represents dependencies between utterances, edge types $\mathcal{R}$ encode temporal and emotional relationships into five categories: self-past, self-future, inter-speaker past, inter-speaker future,

and self-loop. Edge weights $\mathcal{W}$ represent interaction strength or proximity. Adjacency range is constrained by window size $w = 5$ to reduce long-distance noise and computational complexity. During batch processing, dialogue maximum length is set to $L = \max_b N_b$, with padded positions set to zero. The two subgraphs are stacked as $\mathbf{Adj}_s$ and $\mathbf{Adj}_c$, providing structural information for subsequent graph attention modeling.

The intra-speaker subgraph $\mathbf{Adj}_s$ only connects utterances from the same speaker, used to model internal consistency and temporal evolution:

$$(\mathbf{Adj}_s)_{i,j} = \begin{cases} 1 & \text{if } i = j \wedge s_i = s_j, \\ 2 & \text{if } i > j \wedge |i-j| \le w \wedge s_i = s_j \quad \text{(intra-past)}, \\ 3 & \text{if } i < j \wedge |i-j| \le w \wedge s_i = s_j \quad \text{(intra-future)}, \\ 0 & \text{otherwise.} \end{cases} \tag{7}$$

This design highlights the cumulative impact of past utterances on current emotions while introducing forward-looking context from future utterances.

The inter-speaker subgraph $\mathbf{Adj}_c$ is used to model dynamic interactive relationships between different speakers:

$$(\mathbf{Adj}_c)_{i,j} = \begin{cases} 1 & \text{if } i = j, \\ 4 & \text{if } i < j \wedge |i-j| \le w \wedge s_i \ne s_j \quad \text{(inter-future)}, \\ 5 & \text{if } i > j \wedge |i-j| \le w \wedge s_i \ne s_j \quad \text{(inter-past)}, \\ 0 & \text{otherwise.} \end{cases} \tag{8}$$

This subgraph characterizes responses, conflicts, and collaborations in inter-speaker emotional dynamics.

The two relational subgraphs are finally represented as $\mathbf{Adj}_s, \mathbf{Adj}_c \in \mathbb{Z}^{B\times L\times L}$,where $r_{i,j} \in \mathcal{R}$,and $\mathbf{Adj}_{i,j} = \mathrm{id}(r_{i,j})$. The separated subgraph structures support hierarchical modeling, first aggregating global interaction information across speakers, then refining single-speaker internal representations, thereby enhancing modeling capabilities in multi-speaker scenarios.

The above relational subgraphs are shared across different modalities but act on their respective modality representations. Specifically, text, visual, and audio modalities use independent differential graph attention networks and model on the same $\mathbf{Adj}_s$ and $\mathbf{Adj}_c$. Taking text modality as an example, inter-speaker GAT ('gatTer') first acts on $\mathbf{Adj}_c$, followed by intra-speaker GAT ('gatT') on $\mathbf{Adj}_s$. Audio and visual modalities use 'gatAer' / "gatA' and "gatVer' / "gatV' respectively. This design allows each modality to independently learn relation representations while sharing dialogue structures.

4.5. Differential attention graph convolutional network

To fuse structural information in relation subgraphs, we propose the Differential Relational Graph Convolutional Network (DiffRGCN). This network introduces a differential attention mechanism based on GAT to model relation types and feature differences. DiffRGCN consists of multi-head attention layers and an output layer, where each attention head implicitly characterizes enhancement and suppression relationships between nodes through the differential mechanism and enhances semantic awareness by combining relation labels. This network can be applied independently to different modalities and sequentially acts on inter-speaker subgraphs and intra-speaker subgraphs, first aggregating inter-speaker interactions, then refining intra-speaker representations, with the specific process shown in Algorithm 13.

The core of DiffRGCN is the differential graph attention layer. This layer divides the input features into positive and negative branches, modeling emotional enhancement and suppression signals respectively, and achieves differential aggregation through learnable lambda parameters. Specifically, the input features first undergo linear projection:

$$\mathbf{Wh} = \mathbf{H}^m\mathbf{W}, \tag{9}$$

where $\mathbf{H}^m \in \mathbb{R}^{B\times L\times d_{\text{in}}}$, $\mathbf{W} \in \mathbb{R}^{d_{\text{in}}\times d_{\text{out}}}$, B is the batch size, and L is the number of nodes.

Algorithm 1: Differential graph attention layer.

Input: Node features $\mathbf{H}^m$, adjacency matrix $\mathbf{Adj}$, layer depth d, relation-aware flag

Project node features using Eqn. (9) to obtain $\mathbf{Wh}$;
Split $\mathbf{Wh}$ into positive and negative branches;
Generate $\mathbf{q}^{\text{pos}}, \mathbf{k}^{\text{pos}}, \mathbf{v}^{\text{pos}}$ and $\mathbf{q}^{\text{neg}}, \mathbf{k}^{\text{neg}}, \mathbf{v}^{\text{neg}}$ using Eqn. (10) and Eqn. (11);
if *relation-aware enabled* **then**
    Embed adjacency relations into attention scores;
Compute raw attention scores $\mathbf{e}^{\text{pos}}$ and $\mathbf{e}^{\text{neg}}$ using Eqn. (12);
Apply activation, masking, and Softmax normalization to obtain $\boldsymbol{\alpha}^{\text{pos}}$ and $\boldsymbol{\alpha}^{\text{neg}}$;
Compute differential coefficients λ_1, λ_2, and λ_{full} using Eqn. (13) and Eqn. (14);
Fuse positive and negative attention weights to obtain final $\boldsymbol{\alpha}$ using Eqn. (15);
Concatenate value vectors $\mathbf{v}^{\text{pos}}$ and $\mathbf{v}^{\text{neg}}$ to obtain $\mathbf{Wh}_v$ (Eqn. (16));
Aggregate $\mathbf{Wh}_v$ using attention weights $\boldsymbol{\alpha}$ to obtain $\mathbf{h}'$ (Eqn. (17));
Concatenate multi-head outputs to obtain final node representations $\mathbf{x}$ (Eqn. (18));
return $\mathbf{x}$ and attention weights $\boldsymbol{\alpha}$;

In the positive branch, the projected features are used to generate query and key vectors:

$$\mathbf{q}^{\text{pos}} = \mathbf{a}^{\text{pos}}_{\text{left}}\left(\mathbf{Wh}^{\text{pos}}\right), \quad \mathbf{k}^{\text{pos}} = \mathbf{a}^{\text{pos}}_{\text{right}}\left(\mathbf{Wh}^{\text{pos}}\right). \tag{10}$$

The negative branch correspondingly generates $\mathbf{q}^{\text{neg}}$ and $\mathbf{k}^{\text{neg}}$.

The value vectors for positive and negative branches are defined as:

$$\mathbf{v}^{\text{pos}} = \mathbf{Wh}^{\text{pos}}, \quad \mathbf{v}^{\text{neg}} = \mathbf{Wh}^{\text{neg}}. \tag{11}$$

When relation awareness is enabled, the adjacency matrix is embedded as a relation representation $\mathbf{R} \in \mathbb{R}^{B\times L\times L\times d_r}$ and projected to obtain $\mathbf{rel}^{\text{pos}}$ and $\mathbf{rel}^{\text{neg}}$, which are added to the base attention scores:

$$\begin{aligned} \mathbf{e}^{\text{pos}}_{i,j} &= \mathbf{q}^{\text{pos}}_i + \left(\mathbf{k}^{\text{pos}}_j\right)^\top + \mathbf{rel}^{\text{pos}}_{i,j}, \\ \mathbf{e}^{\text{neg}}_{i,j} &= \mathbf{q}^{\text{neg}}_i + \left(\mathbf{k}^{\text{neg}}_j\right)^\top + \mathbf{rel}^{\text{neg}}_{i,j}. \end{aligned} \tag{12}$$

The positive attention scores undergo LeakyReLU activation, mask invalid edges with $\mathbf{Adj} \le 0$, and are then normalized via Softmax to obtain positive attention weights $\boldsymbol{\alpha}^{\text{pos}}$. The negative branch follows the same process to obtain $\boldsymbol{\alpha}^{\text{neg}}$.

To balance positive and negative attention, differential lambda parameters are introduced. First, we compute:

$$\begin{aligned} \lambda_1 &= \exp\left(\sum_{d=1}^{D} \lambda^{(d)}_{\text{left},1} \cdot \lambda^{(d)}_{\text{right},1}\right), \\ \lambda_2 &= \exp\left(\sum_{d=1}^{D} \lambda^{(d)}_{\text{left},2} \cdot \lambda^{(d)}_{\text{right},2}\right), \end{aligned} \tag{13}$$

where all parameters are initialized from a normal distribution with mean 0 and variance 0.1. The final differential coefficient is defined as:

$$\lambda_{\text{full}} = \lambda_1 - \lambda_2 + \lambda_{\text{init}}, \tag{14}$$

where $\lambda_{\text{init}} = 0.8 - 0.6\exp(-0.3 \cdot \text{depth})$, and depth denotes the layer depth.

The final attention weights are obtained through differential fusion:

$$\boldsymbol{\alpha} = \boldsymbol{\alpha}^{\text{pos}} - \lambda_{\text{full}} \cdot \boldsymbol{\alpha}^{\text{neg}}. \tag{15}$$

This operation enhances effective emotional dependencies while suppressing noise interference.

Subsequently, the value vectors of the positive and negative branches are concatenated:

$$\mathbf{Wh}_v = \text{cat}(\mathbf{v}^{\text{pos}}, \mathbf{v}^{\text{neg}}) \in \mathbb{R}^{B\times L\times d_{\text{out}}}, \tag{16}$$

and neighbor feature aggregation is completed under the attention weights:

$$\mathbf{h}' = \alpha \mathbf{Wh}_v. \tag{17}$$

DiffRGCN adopts multi-head differential attention to model multi-view relationships in parallel, with outputs concatenated as:

$$\mathbf{x} = \text{cat}\big([\text{att}_i(\mathbf{x}, \mathbf{Adj})]\big). \tag{18}$$

The concatenated result undergoes dropout, an output attention layer, and a fully connected layer with residual connection, and final node representations are obtained via layer normalization.

This design combines multi-head mechanisms with differential modeling, effectively enhancing the expressive power and stability of relation modeling while keeping the computational complexity $O(BLwd \cdot h + BL^2/h)$ controllable.

4.6. Adaptive modality balancing

In multimodal learning, the contribution of each modality to emotion modeling is typically imbalanced, which may cause weak modalities to be overlooked during fusion. To address this issue, we propose a Dynamic Modality Balancing mechanism, which dynamically adjusts modality weights via an Adaptive Modality Dropout strategy. The detailed procedure is illustrated in Algorithm 15. To ensure timely responsiveness to changes in modality contributions, the modality performance metric p_m, ratio parameter $r_{m,j}$, and dropout probability q_m are updated in each training batch. This mechanism is only activated after the model enters a stable training phase. The determination of the stable training phase is based on the sensitivity experiment results of the warm-up epochs. As shown in the experiments in Section 6.2 the model performance stabilizes and reaches the optimal level after approximately 60 training epochs on the IEMOCAP dataset. Therefore, this paper sets the warm-up epoch to about 60 to avoid the impact of early parameter fluctuations on model convergence.

First, we calculate the modality-level performance metric p_m for each instance. We use the weighted F1-score for measurement: $p_m = \text{F1}(\hat{y}_m, y, \mathbf{U})$, where $\hat{y}_m = \arg\max(\hat{\mathbf{P}})$, y denotes the ground-truth label, and $\mathbf{U}$ is the utterance mask used to filter padding positions. It is defined as:

$$p_m = \frac{\sum_{c=1}^{C} w_c \cdot \text{Precision}_c \cdot \text{Recall}_c}{\sum_{c=1}^{C} w_c}, \tag{19}$$

where $w_c = \frac{N}{\sum_i \mathbb{I}(y_i=c)}$ is the inverse frequency weight of class c. Precision and Recall are computed based on valid utterances. Specifically, we first flatten the logits and labels: $\mathbf{P}_m^{\flat} = \mathbf{P}_m[\mathbf{U} == 1]$, $y^{\flat} = y[\mathbf{U} == 1]$, then calculate $\text{Precision}_c = \frac{\text{TP}_c}{\text{TP}_c+\text{FP}_c}$, $\text{Recall}_c = \frac{\text{TP}_c}{\text{TP}_c+\text{FN}_c}$.

Subsequently, we compute the relative performance differences between modalities. For modality m, its performance ratio with other modalities is defined as:

$$r_{m,j} = \frac{p_m}{p_j + \epsilon} - 1, \quad \forall j \neq m, \tag{20}$$

where $\epsilon = 10^{-5}$ is a smoothing factor to avoid division by zero. Furthermore, for any modality m, the dimension of its relative performance ratio vector $\mathbf{r}_m$ is $|\mathcal{M}| - 1$, where $\mathcal{M} = \{t, v, a\}$. Each dimension of $\mathbf{r}_m$ corresponds to the performance ratio $r_{m,j}$ between modality m and another modality $j (j \neq m)$, i.e.:

$$\mathbf{r}_m = [r_{m,j}]_{j\in\mathcal{M}, j\neq m}. \tag{21}$$

In subsequent calculations, we do not explicitly filter positive values of $\mathbf{r}_m$; instead, we directly apply the ReLU activation to the complete ratio vector to suppress negative terms, and perform Softmax normalization on the non-negative results to smooth the relative performance differences between modalities. Specifically, we construct the ratio vector $\mathbf{r}_m$ by arranging all relative performance ratios $r_{m,j} (j \in \mathcal{M}, j \neq m)$ in index order, first applying non-negative constraint: $\mathbf{r}_m^+ = \max(\mathbf{r}_m, 0)$, then performing Softmax normalization:

$$\hat{\mathbf{r}}_m = \frac{\exp(\mathbf{r}_m^+)}{\sum \exp(\mathbf{r}_m^+)}, \tag{22}$$

and finally obtaining the weighted average ratio:

$$\bar{r}_m = \frac{\sum \hat{\mathbf{r}}_m \odot \mathbf{r}_m}{|\mathbf{r}_m|}. \tag{23}$$

Based on this ratio, the modality dropout probability is defined as:

$$q_m = q_{\text{base}} \cdot (1 + \lambda \cdot \bar{r}_m), \quad q_m = \text{clip}(q_m, 0, 1), \tag{24}$$

where $q_{\text{base}} = 0.3$ and $\lambda = 0.9$ are hyperparameters. This design makes high-performance modalities have a higher dropout probability, thereby encouraging the model to focus on weaker modalities while avoiding extreme differences from dominating the training process.

When performing modality dropout, for the modality feature set $\mathbf{F} = [\mathbf{T}, \mathbf{V}, \mathbf{A}] \in \mathbb{R}^{M\times B\times N\times d_h}$ (where $M = 3$ and B is the number of instances), we first decide whether to perform dropout with probability $p_{\text{exe}} = 0.5$. If dropout is performed, we generate a Bernoulli mask:

$$\mathbf{M}_{m,b} \sim \text{Bernoulli}(1 - q_m), \tag{25}$$

and apply it to the features: $\mathbf{F}' = \mathbf{M} \odot \mathbf{F}$.

Subsequently, we perform expectation compensation scaling via a custom Autograd function:

$$\mathbf{F}'' = \frac{\mathbf{F}'}{1-\theta}, \quad \theta = \frac{\sum_m d_m q_m}{\sum_m d_m}, \tag{26}$$

where $d_m = d_h$ is the modality dimension. Its forward propagation is linear scaling, and the backward propagation only passes gradients to the input:

$$\frac{\partial \mathbf{F}''}{\partial \mathbf{F}'} = \frac{1}{1-\theta}, \quad \frac{\partial \mathbf{F}''}{\partial \theta} = 0. \tag{27}$$

Algorithm 2: Adaptive modality dropout.

Input: Modality feature list $\mathbf{F} = [\mathbf{T}, \mathbf{V}, \mathbf{A}]$, batch size B, base dropout probability q_{base}, scaling factor λ, labels $\mathbf{y}$, training mask $\mathbf{U}$

for *each training batch* **do**
  `// Compute modality performance` p_m
  for *each modality* m **do**
    $p_m = \text{F1}(\mathbf{F}_m, \mathbf{y}, \mathbf{U})$;
  `// Compute relative performance ratios and dropout probabilities`
  for *each modality* m **do**
    $r_{m,j} = \frac{p_m}{p_j+\epsilon} - 1, \forall j \neq m$;
    $q_m = \text{clip}(q_{\text{base}} \cdot (1 + \lambda \cdot \text{Softmax}(\text{ReLU}(r_{m,j}))), 0, 1)$;
  `// Apply dropout and scaling`
  Randomly decide whether to apply dropout;
  if *dropout is applied* **then**
    Sample Bernoulli mask $\mathbf{M} \sim 1 - q$;
    Apply mask: $\mathbf{F}'' = \mathbf{M} \odot \mathbf{F}$;
    Scale by expectation: $\mathbf{F}'' = \mathbf{F}''/(1-\theta)$;
    Keep instances with at least one modality: $\mathbf{F}''' = \mathbf{F}''[\mathbf{u}]$;
  else
    $\mathbf{F}''' = \mathbf{F}, \mathbf{u} = \mathbf{1}$;
return Dropped and scaled modality features $\mathbf{F}'''$ and valid instance mask $\mathbf{u}$.

Finally, we filter valid instances using the update flag $\mathbf{u}_b = \sum_m \mathbf{M}_{m,b} > 0$, retaining only samples with at least one modality: $\mathbf{F}''' = \mathbf{F}''[\mathbf{u}]$, thus avoiding the empty representation problem caused by full modality dropout.

4.7. Emotion classifier

After obtaining multimodal dynamic representations, we design a multimodal fusion and classification module to integrate text, visual, and audio features and generate emotion predictions. For the final representation of each modality, the model uses independent classification heads for projection. Each classification head consists of ReLU activation, Dropout layer, and linear mapping to map hidden representations to class space. Taking the text modality as an example, its classification process is defined as:

$$\mathbf{t} = \sigma(\text{Dropout}(\text{ReLU}(\mathbf{x}_i^t \mathbf{W}_t + \mathbf{b}_t))). \tag{28}$$

where $\mathbf{W}_t \in \mathbb{R}^{d_h \times C}$, $\mathbf{b}_t \in \mathbb{R}^C$, σ represents the output after linear mapping. The classification heads $\mathbf{v}$ and $\mathbf{a}$ for visual and audio modalities are constructed in the same way.

Subsequently, fusion is completed by element-wise addition of each modality's logits:

$$\mathbf{O} = \mathbf{t} + \mathbf{v} + \mathbf{a}, \tag{29}$$

and log-softmax is used to compute the fused probability distribution:

$$\mathbf{P} = \log(\text{softmax}(\mathbf{O})), \quad \hat{\mathbf{P}} = \text{softmax}(\mathbf{O}). \tag{30}$$

The predicted label is obtained by argmax of $\hat{\mathbf{P}}$. To enhance unimodal representation capabilities, the model simultaneously computes unimodal log-softmax probabilities $\mathbf{P}_t$, $\mathbf{P}_v$ and $\mathbf{P}_a$ for auxiliary optimization.

The model training uses cross-entropy loss and incorporates sequence masks to adapt to variable-length inputs. The fusion loss is defined as:

$$\mathcal{L}_{\text{fusion}} = \text{CE}(\mathbf{P}, \mathbf{y}, \mathbf{m}), \tag{31}$$

where the cross-entropy form is:

$$\text{CE}(\mathbf{P}, \mathbf{y}, \mathbf{m}) = -\sum_{i=1}^{L} m_i \sum_{c=1}^{C} y_{i,c} \log P_{i,c}. \tag{32}$$

Here, L represents sequence length, C represents number of classes, $P_{i,c}$ is the predicted probability, $y_{i,c}$ is the one-hot label, $m_i \in 0, 1$ is the mask. The final training objective consists of fusion loss and unimodal auxiliary losses:

$$\mathcal{L} = \mathcal{L}_{\text{fusion}} + \alpha_t \mathcal{L}_t + \alpha_a \mathcal{L}_a + \alpha_v \mathcal{L}_v, \tag{33}$$

where $\alpha_t = \mathcal{L}_t / 10$ etc. are adaptive weights to balance training contributions from different modalities.

5. Experiments

This section introduces the experimental setup of this study, including datasets, baseline models, evaluation metrics, and implementation details. Through these settings, we conducted a comprehensive evaluation of the proposed model and compared it with existing state-of-the-art methods to validate its effectiveness.

5.1. Datasets

As shown in Table 1, this experiment employs two widely used multimodal dialogue emotion recognition datasets: IEMOCAP and MELD. These datasets contain multimodal features such as text, audio, and vision, making them suitable for evaluating the model's performance in multimodal fusion and dialogue context modeling.

IEMOCAP dataset (Busso et al., 2008): This dataset was collected by the University of Southern California, consisting of 10 dialogue sessions with a total of approximately 12 hours of video recordings. Each dialogue involves two speakers and is annotated with 6 emotion categories: neutral, happy, sad, angry, excited, and frustrated. The dataset comprises a total of 5531 utterances. We follow the standard partitioning method, using the first 8 sessions as the training set and the last 2 as the test set. The multimodal nature of this dataset makes it a benchmark for evaluating emotion recognition models.

MELD dataset (Poria et al., 2018): This dataset is based on dialogue segments from the TV show "Friends," containing approximately 1400 dialogues with 13,708 utterances. It is annotated with 7 emotion categories: neutral, surprise, fear, sad, joy, disgust, and angry. We use the official split: approximately 1000 dialogues for the training set, 100 for the validation set, and 300 for the test set. The larger number of speakers in this dataset (up to 9) increases the complexity of dialogue context modeling.

5.2. Baselines

To validate the superiority of the proposed model, we selected the following baseline models for comparison. These models cover representative methods based on RNN, GCN, and multimodal fusion. We reproduced them on the same datasets or used publicly available code for experiments.

DialogueRNN (Majumder et al., 2019): An RNN-based model that utilizes attention mechanisms to capture emotional dynamics in dialogues.

DialogueGCN (Ghosal et al., 2019): A graph convolutional network-based model that models dialogues as graph structures to capture dependencies between utterances.

MMGCN (Hu et al., 2021): A multimodal graph attention network that fuses text, audio, and visual modalities through graph attention mechanisms for emotion recognition.

MM-DFN (Hu et al., 2022): A multimodal dynamic fusion network that uses dynamic attention to fuse multimodal features.

COGMEN (Joshi et al., 2022): A contextualized graph neural network for multimodal emotion recognition, combining GNN and multimodal fusion.

MultiEMO (Shi & Huang, 2023): An attention-based correlation-aware multimodal fusion framework that emphasizes inter-modal correlations.

SDT (Ma et al., 2024): A speaker-dependent Transformer model that uses Transformers to capture speaker-specific patterns.

GraphCFC (Li et al., 2023): A directed graph-based cross-modal fusion network for multimodal emotion analysis.

RL-EMO (Zhang et al., 2024): A reinforcement learning-enhanced emotion recognition model that optimizes emotion prediction via RL.

DEDNet (Wang et al., 2024): A dual encoder-decoder network for handling multimodal dialogue emotions.

DER-GCN (Ai et al., 2024): The model constructs a weighted multi-relational graph to capture diverse dependencies in dialogues and integrates multimodal features via a self-supervised masked graph autoencoder and a multi-information Transformer.

MERC-PLTAF (Wu et al., 2025b): This work adopts fine-grained feature extraction and cross-modal fusion strategies to jointly model multimodal emotional information across dialogues.

5.3. Evaluation metrics

We employ the weighted F1 score and weighted accuracy as the primary evaluation metrics to comprehensively assess model performance. All metrics are computed on the test set using the Scikit-learn library.

5.4. Implementation details

The model is implemented using the PyTorch framework. For the text modality, features are extracted using RoBERTa with a dimension

Table 1
Statistical information of the IEMOCAP and MELD datasets, including the number of conversations, utterances, speakers, total emotion classes, and the number of utterances for each emotion class. The values are reported as Train + Validation / Test splits.

Dataset	Convs	Utterances	Speakers	Classes	Neutral	Happy Joy	Sadness	Angry	Excited Surprise	Frustrated	Disgust	Fear
IEMOCAP	120 / 31	5810 / 1623	2	6	1708 / 490	1636 / 456	1084 / 266	1103 / 290	1041 / 265	1848 / 324	-	-
MELD	1153 / 280	11,098 / 2610	9	7	5180 / 1256	1940 / 368	794 / 208	1243 / 364	1205 / 431	-	293 / 68	268 / 50

Table 2
The table evaluates the performance of all models on the IEMOCAP (six emotion categories) datasets using F1 scores, while presenting their overall performance across three datasets with weighted accuracy (wa-ACC) and weighted F1 score (wa-F1) as metrics, where the best results are bolded and the second-best are underlined.

Models	IEMOCAP													
	happy		sad		neutral		angry		excited		frustrated		wa-ACC	wa-F1
	ACC	F1	ACC	F1	ACC	F1	ACC	F1	ACC	F1	ACC	F1		
DialogueRNN	25.00	34.95	82.86	84.58	54.43	57.66	61.76	64.42	**90.97**	76.30	62.20	59.55	65.43	64.29
DialogueGCN	64.29	29.03	80.86	64.37	43.14	50.96	68.49	63.29	71.85	68.19	57.68	62.41	62.07	58.19
MMGCN	32.64	39.66	72.65	76.89	65.10	62.81	73.53	71.43	77.93	75.40	65.09	63.43	66.62	66.25
MM-DFN	44.44	44.44	77.55	80.00	71.35	66.99	75.88	70.88	74.25	76.42	58.27	61.67	67.84	67.85
TS-GCL	71.20	70.00	81.30	81.70	67.40	64.20	60.50	61.40	74.60	76.50	62.00	64.00	70.30	70.20
MultiEMO	53.80	56.29	83.33	83.50	**75.60**	70.11	68.29	67.07	79.70	75.79	64.82	70.35	72.29	71.69
SDT	55.06	57.62	80.58	80.08	65.73	69.14	67.88	66.87	82.50	73.47	66.58	67.53	70.54	69.95
GraphCFC	41.52	45.08	**87.12**	**84.84**	65.19	63.27	68.31	70.82	77.16	75.85	62.86	63.19	68.39	68.02
RL-EMO	40.28	47.15	79.18	81.17	69.79	66.67	74.12	64.28	78.60	76.18	58.01	61.82	69.16	68.20
DEDNet	56.32	64.07	81.15	80.98	73.92	74.97	67.37	71.11	84.38	77.84	72.99	69.68	74.47	73.79
DER-GCN	60.70	58.80	75.90	79.80	66.50	61.50	71.30	72.10	71.10	73.30	66.10	67.80	69.70	69.40
MERC-PLTAF	**75.60**	**75.30**	74.20	80.00	71.10	71.40	**75.90**	**74.70**	59.70	54.70	**73.50**	**72.20**	72.70	71.40
AMB-DSGDN(Ours)	60.23	66.25	84.58	81.36	74.14	**76.20**	70.86	71.88	85.87	**81.34**	73.24	72.17	**76.09**	**75.64**

Table 3
The table evaluates the performance of all models on the MELD (seven emotion categories) datasets using F1 scores, while presenting their overall performance across three datasets with wa-ACC and wa-F1 as metrics, where the best results are bolded and the second-best are underlined.

Models	MELD															
	neutral		surprise		fear		sadness		joy		disgust		anger		wa-ACC	wa-F1
	ACC	F1	ACC	F1	ACC	F1	ACC	F1	ACC	F1	ACC	F1	ACC	F1		
DialogueRNN	82.17	76.56	46.62	47.64	0.00	0.00	21.15	24.65	49.50	51.49	0.00	0.00	48.41	46.01	60.27	57.95
MMGCN	84.32	76.96	47.33	49.63	2.00	3.64	14.90	20.39	56.97	53.76	1.47	2.82	42.61	45.23	61.34	58.41
MM-DFN	79.06	75.80	53.02	50.42	0.00	0.00	17.79	23.72	59.20	55.48	0.00	0.00	50.43	48.27	60.96	58.72
TS-GCL	78.10	**80.60**	56.70	56.40	6.80	5.20	42.30	43.70	**68.30**	66.30	2.30	2.60	43.80	48.50	64.40	64.10
MultiEMO	76.44	79.42	55.92	58.12	25.71	21.18	51.05	41.60	62.23	63.07	**45.71**	31.07	55.28	53.37	65.45	65.77
SDT	75.99	79.65	**59.21**	58.78	24.44	23.16	**59.22**	39.23	64.40	62.76	40.00	**31.86**	52.27	54.44	66.00	65.92
GraphCFC	71.26	75.17	46.18	45.68	9.09	3.28	30.43	11.42	50.26	49.49	0.00	0.00	35.88	41.93	54.34	55.20
RL-EMO	**85.59**	79.57	58.72	59.03	14.00	16.09	18.27	27.64	60.45	63.53	16.18	20.18	**55.36**	52.84	65.63	63.47
DEDNet	76.51	79.97	56.77	58.90	21.88	17.07	50.38	39.30	64.97	62.63	40.54	28.57	53.65	**54.49**	65.52	65.88
MERC-PLTAF	82.90	80.50	59.10	59.10	24.00	**26.90**	55.30	**46.40**	63.40	**77.10**	17.60	26.10	50.30	53.90	**68.00**	52.80
AMB-DSGDN(Ours)	75.16	80.00	58.48	**59.30**	**30.00**	17.14	58.18	40.25	64.75	64.59	44.44	25.26	54.93	54.12	66.07	**66.18**

of 1024. For the visual modality, DenseNet is employed to extract features with a dimension of 342; for the audio modality, OpenSmile is used to extract features with a dimension of 1582 for the IEMOCAP dataset or 300 for the MELD dataset. These initial features are projected to a unified hidden dimension of 512 through a linear layer for subsequent processing. To ensure training stability and model performance, the batch size is set to 16, the learning rate is 0.000068, the Adam optimizer is used with a weight decay of 0.00005, and training is conducted for 100 epochs. The modality dropout mechanism has a base probability $q_{\text{base}} = 0.3$ and an execution probability $p_{\text{exe}} = 0.5$.

6. Results and analysis

This section systematically evaluates the proposed AMB-DSGDN model from multiple perspectives, including overall performance, ablation studies, parameter sensitivity, robustness, and computational complexity, to comprehensively analyze its effectiveness and practicality.

6.1. Main results

Tables 2 and 3 show the performance of the proposed model AMB-DSGDN on the IEMOCAP and MELD datasets, compared with various baseline models. To ensure experimental fairness, all models use the same preprocessing and training settings.

On the IEMOCAP dataset, AMB-DSGDN achieved a wa-ACC of 76.09% and a wa-F1 of 75.64%, improving by 1.62% and 1.85% respectively compared to the second-best model DEDNet. Especially on anger, excitement, and frustration emotions, the model shows significant advantages, benefiting from the differential attention graph convolutional network's ability in fine-grained semantic relation modeling, as well as the adaptive modality dropout strategy's effective alleviation of modality imbalance, enabling the model to better capture key emotional information in dialogues.

On the MELD dataset, AMB-DSGDN's weighted accuracy is 66.07%, and weighted F1 score is 66.18%, with limited improvements compared to the second-best model. We believe there are multiple factors

Table 4
Comparison of performance of different modal combinations on IEMOCAP and MELD datasets.

Modality Setting	IEMOCAP		MELD	
	wa-ACC	wa-F1	wa-ACC	wa-F1
A	62.60	60.70	34.70	37.70
T	69.02	68.63	65.61	65.78
V	39.80	35.60	23.16	31.27
A + T	73.43	72.87	65.49	65.65
V + T	69.71	69.16	65.77	65.86
A + V + T	**76.09**	**75.64**	**66.07**	**66.18**

contributing to this. First, as shown in Table 1, MELD's class distribution is highly imbalanced, with sparse samples for several emotions (such as disgust, fear), limiting the model's learning and generalization on low-frequency categories; Second, MELD is multi-speaker dialogue, where audio and visual signals are more susceptible to interference from speaker switching, overlapping speech, facial occlusion, etc., thereby reducing unimodal quality and increasing modality fusion difficulty; Nevertheless, the model still shows obvious advantages on categories like surprise, demonstrating its adaptability in multi-speaker environments.

Overall, AMB-DSGDN shows significant advantages on IEMOCAP due to higher modality quality and clearer emotional associations; while on MELD, influenced by class distribution imbalance and modality quality fluctuations caused by multi-speakers, overall improvements are limited, but it still has advantages on categories like surprise, reflecting the model's robustness and adaptability.

6.2. Ablation study

To thoroughly validate the effectiveness of key components in the model, we conduct a series of ablation experiments, focusing on factors such as modality combinations, window sizes, differential attention mechanisms, and adaptive modality dropout strategies. By progressively removing or adjusting these modules and retraining/evaluating on the IEMOCAP and MELD datasets, we quantify the contributions of each component and their synergistic effects.

Multimodal combinations: In Table 4, we examine the impact of different modality combinations on model performance to verify the effectiveness of multimodal fusion. Specifically, we evaluate performance under unimodal, bimodal, and full-modal settings and compare across the two datasets. The results show that in unimodal configurations, the text modality generally achieves the best performance, reflecting its advantage in capturing dialogue semantics and contextual information, followed by the audio modality, likely due to its sensitivity to speech tone and rhythm. In contrast, the visual modality performs relatively weakly, possibly because visual cues (such as facial expressions) are limited by perspectives and dynamic changes in dialogue scenes, leading to instability in information extraction. Furthermore, in bimodal settings, the combination of audio and text significantly outperforms the visual-text combination, indicating that the audio modality better complements textual information by providing additional acoustic cues to enhance emotion recognition accuracy. While the visual-text fusion shows improvements, the gains are limited, suggesting that the visual modality may introduce noise or redundancy in certain dialogue contexts. Ultimately, full-modal integration (i.e., combining audio, visual, and text) achieves the best performance on both datasets, validating the importance of multimodal synergy. By leveraging multiple information sources simultaneously, the model can more comprehensively capture emotional dynamics, alleviating the limitations of unimodal or bimodal approaches and thereby improving overall robustness and generalization.

Window size settings: To explore the impact of contextual range on graph convolutional networks in dialogue emotion recognition tasks,

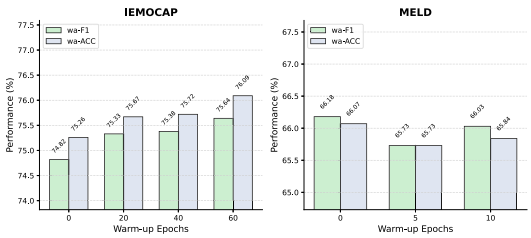

Fig. 3. Experimental results of different window sizes on IEMOCAP and MELD datasets. The purple line represents wa-ACC, the light yellow bar represents wa-F1; the left subplot corresponds to IEMOCAP dataset, the right subplot to MELD dataset. Note: Window size determines the semantic association capture range of the graph convolutional network and needs to be adjusted based on dataset characteristics. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 5
Ablation study of semantic graph differential network on IEMOCAP and MELD datasets.

Variants	IEMOCAP		MELD	
	wa-F1	wa-ACC	wa-F1	wa-ACC
w/o Rel Subgraph + DiffRGCN	68.00	68.82	65.07	65.10
w/o DiffRGCN	73.44	73.56	65.62	64.80
Full Model	**75.64**	**76.09**	**66.07**	**66.18**

this experiment systematically evaluates the role of different window sizes on model performance. Fig. 3 shows that on the IEMOCAP dataset, as window size gradually increases, the model's weighted average accuracy and F1 values reach optimal performance under medium window configurations. This indicates that moderate windows can fully cover semantically related utterances and their contextual dependencies, thereby improving the graph structure's modeling ability for emotional relationships. Too small windows limit the capture of long-range dependency information, leading to insufficient semantic information; too large windows may introduce irrelevant features, accumulating noise, thereby reducing discriminative ability in attention weight calculation and feature propagation. On the MELD dataset, optimal performance is concentrated in relatively smaller window intervals. Considering that MELD's dialogues have more frequent inter-speaker interactions and more complex multi-turn structures, smaller windows can focus more on local semantics and multimodal information interactions, while overly large windows easily incorporate irrelevant emotional features across speakers, increasing modality inconsistencies and distribution noise. This shows that different datasets' dialogue structures, speaker distributions, and emotional interaction patterns significantly affect the optimal choice of window size, and in practical applications, this hyperparameter needs to be tuned based on data statistical characteristics.

Semantic graph differential network: To verify the role of DiffRGCN and relational subgraph components in the overall model, we designed targeted ablation studies by replacing or removing related modules while keeping the rest of the structure consistent to analyze the independent contributions of each component to performance. It should be noted that the adaptive modality balancing mechanism remains enabled in this group of ablation studies. Specifically, when only DiffRGCN is removed, we replace the differential relational graph convolution with a regular graph convolutional neural network for feature propagation on the constructed relational subgraphs, with the rest of the model architecture unchanged; when both relational subgraphs and DiffRGCN are removed, the model no longer performs graph structure modeling, but directly connects the adaptive modality balancing mechanism and subsequent classification modules after the utterance-level multimodal encoder. The results in Table 5 show that simultaneously re-

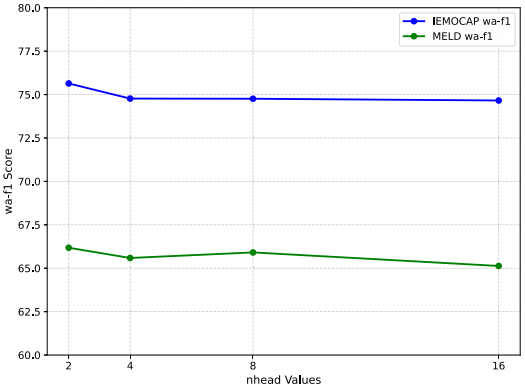

Fig. 4. Sensitivity analysis results of attention head numbers on IEMOCAP and MELD datasets. The blue line represents wa-F1 on IEMOCAP dataset, the green line represents wa-F1 on MELD dataset, horizontal axis "nhead Values" indicates number of attention heads, vertical axis "wa-F1 Score" indicates corresponding weighted average F1 score. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 6
Ablation study of modality balancing on IEMOCAP and MELD datasets.

Variants	IEMOCAP		MELD	
	wa-F1	wa-ACC	wa-F1	wa-ACC
w/o MD + Rel Subgraph + DiffRGCN	67.83	68.78	64.87	64.80
w/o MD + DiffRGCN	72.99	73.56	65.56	65.71
w/o MD	74.77	75.29	66.00	66.05
Full Model	**75.64**	**76.09**	**66.07**	**66.18**

moving relational subgraphs and DiffRGCN causes significant declines in weighted F1 and accuracy on IEMOCAP and MELD, verifying the importance of semantic relation modeling. Further comparison reveals that replacing DiffRGCN with GCN alone leads to performance degradation but still outperforms the variant without graph structures entirely, indicating that relational subgraphs provide basic semantic associations, while DiffRGCN utilizes differential attention to characterize fine-grained semantic differences between nodes, effectively filtering redundant information and strengthening the representation of multimodal emotional dependencies. The above results fully prove the key value of the semantic graph differential network in multimodal learning and its positive impact on model performance and generalization capabilities.

Sensitivity of attention heads: In further sensitivity analysis, we explored the impact of the number of attention heads in the differential attention graph convolutional network on model performance to reveal its role in multimodal feature aggregation. As shown in Fig. 4, on the IEMOCAP dataset, the model performs best under a small number of attention heads. This is because a small number of attention heads can effectively capture semantic diversity while keeping feature representations compact. Too few heads may lead to insufficient multimodal information integration, while too many heads may introduce redundant features or increase training difficulty, thereby reducing model discriminative ability. On the MELD dataset, the model also performs excellently under low head configurations, with performance slightly declining as the number of heads increases. Attributed to the multi-speaker dialogue characteristics of the MELD dataset, the integration difficulty of cross-modal signals is higher in multi-speaker environments, too many attention heads may introduce unnecessary noise and information conflicts, while a streamlined attention mechanism can more concentratedly capture core semantic associations and emotional dynamics, achieving efficient information integration.

Adaptive modality balancing: As shown in Table 6, we conducted ablation analysis on the synergistic effects of the adaptive modality dropout (MD) strategy and semantic graph components (relational subgraphs and DiffRGCN) by progressively removing core modules. In this group of experiments, the adaptive modality balancing mechanism is disabled by default, with the rest of the network structure and experimental settings consistent with the semantic graph differential network subsection. Specifically, when simultaneously removing the modality dropout mechanism, relational subgraphs, and DiffRGCN, the model no longer performs modality regulation or graph structure modeling, but directly inputs utterance-level multimodal representations into the classification module; when removing the modality dropout mechanism and DiffRGCN, the model only retains relational subgraphs for feature aggregation; when only removing the modality dropout mechanism, the model retains the complete semantic graph structure, but each modality participates in fusion with fixed weights. The experimental results show that simultaneously removing all three leads to the maximum decline in weighted F1 and accuracy on IEMOCAP and MELD, indicating significant complementarity between modality regulation and semantic relation modeling. Further comparison reveals that the model without the modality dropout mechanism still experiences performance decline in multimodal feature fusion, showing that this mechanism can dynamically adjust each modality's participation, suppressing excessive influence from dominant modalities, thereby improving emotion recognition accuracy under noise interference. In summary, the adaptive modality dropout strategy and semantic graph components have significant complementary effects in multimodal dialogue emotion recognition, synergistically enhancing the model's ability to capture emotional dynamic dependencies and robustness to noise interference.

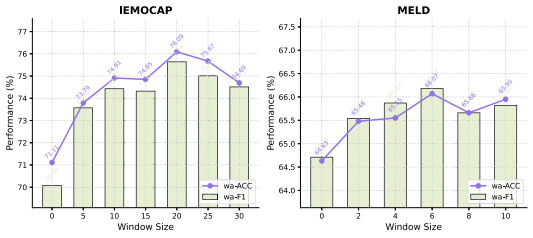

Fig. 5. Performance dynamics in the warm-up phase. The horizontal axis represents the warm-up period (i.e., the number of training rounds before the modality dropout mechanism is gradually activated), the vertical axis is model performance metrics; different color curves correspond to weighted average F1 scores and accuracies on IEMOCAP and MELD datasets respectively, showing the impact of different warm-up periods on model performance.

Performance dynamics during warm-up period: To verify the effectiveness of the adaptive modality dropout strategy, we examined the impact of the warm-up period (i.e., the number of training rounds before the modality dropout mechanism is gradually activated) on model performance to reveal its role in model stability and optimization effects in the early training stage. As shown in Fig. 5, on the IEMOCAP dataset, as the warm-up period extends, the model's performance first rises rapidly, reaching a peak at about 60 training rounds, then remains stable. This indicates that moderate warm-up can help the model gradually adapt to modality imbalance issues, reducing noise interference in early training, thereby improving the capture accuracy of emotional dynamic features. If the warm-up period is too short, the model may enter the modality dropout phase prematurely, leading to insufficient learning of internal representations and cross-modal relationships in each modality's features, resulting in performance fluctuations or unstable convergence. For the MELD dataset, the model's initial performance is higher without warm-up, with performance slowly declining as warm-up rounds increase, and optimal performance appears in shorter warm-up period configurations. Due to the inherent complexity of multi-speaker dialogues, the integration difficulty of cross-modal signals increases accordingly, overly long warm-up periods may cause the model to overly rely on partial modality features without sufficiently focusing on core emotional cues, thereby introducing redundant information interference.

Table 7
Sensitivity analysis results of modal discard probability parameter.

IEMOCAP			
q_{base}	$p_{exe} = 0.3$	$p_{exe} = 0.5$	$p_{exe} = 0.7$
0.1	74.88	75.62	75.2
0.3	75.37	**75.64**	75.06
0.5	75.07	74.96	75.12
MELD			
q_{base}	$p_{exe} = 0.1$	$p_{exe} = 0.2$	$p_{exe} = 0.3$
0.1	65.91	65.8	65.78
0.2	65.80	**66.18**	65.81
0.3	65.87	65.87	65.72

Conversely, shorter warm-up periods can quickly activate the modality dropout mechanism, enabling the model to more effectively identify and weaken redundant features, thereby efficiently integrating cross-modal information, improving training convergence speed, and stabilizing performance.

Dropout probability settings: In the parameter sensitivity analysis of the adaptive modality dropout strategy, we systematically explored the impact of different combinations of base dropout probability q_base and execution probability p_exe on model performance through grid search on two datasets to evaluate its optimization effects. As shown in Table 7, on the IEMOCAP dataset, model performance gradually improves under lower base dropout probability configurations, reaching a peak under medium base dropout probability (e.g., 0.3) and medium execution probability (e.g., 0.5) combinations, then stabilizing. This indicates that moderate dropout can balance noise suppression and key information retention, too low dropout may lead to noise features interfering with fusion results, while too high dropout disrupts semantic representations and increases training instability. Similarly, on the MELD dataset, the model has a higher starting performance under lower base dropout probabilities, with optimal configurations concentrated in relatively lower base dropout probabilities (e.g., 0.2) and medium execution probability intervals, closely related to the dataset's multi-speaker, rich dialogue content, and diverse emotion characteristics. Conservative dropout strategies can effectively retain core information from each modality, reducing interference caused by signal differences between different speakers, while balancing noise suppression, making the model more robust in complex and diverse dialogue environments.

6.3. Handling long-sequence dialogues

This section conducts a dedicated evaluation of the model's capability in long-sequence dialogue scenarios, focusing on its ability to capture dynamic contextual dependencies. The experiments are based on a long-dialogue subset of the IEMOCAP dataset, selecting dialogue samples with 20 to 50 utterances to simulate the complex and evolving temporal dependencies found in real interactive scenarios. Model performance is assessed using weighted F1 score and accuracy, and compared against multiple baseline methods.

The results shown in Fig. 6 indicate that the proposed method significantly outperforms the baseline models on the long-sequence dialogue subset, demonstrating its effectiveness in modeling long-range contextual dependencies. Compared with the baselines, our model maintains stable performance even on longer dialogue sequences, without noticeable degradation, indicating stronger capability in capturing long-term dependencies.

Further analysis reveals that this advantage primarily arises from two design aspects. First, the dynamic cross-modal contribution balancing mechanism adjusts the participation of different modalities in real time according to the dialogue progression, preventing a single modality from dominating and causing information bias over long sequences. Second, the differential relation graph modeling explicitly captures fine-grained differences between nodes, effectively mitigating the common issue of context forgetting in long-sequence tasks. In the middle stages of dialogue, baseline methods often exhibit performance fluctuations due to accumulated noise and modality shifts, whereas the proposed method maintains stable emotion recognition by adaptively adjusting modality weights.

In the later stages of dialogue, our method makes more effective use of accumulated contextual information, suppressing the influence of early noise on current predictions and reducing error propagation. This indicates that the model is capable not only of capturing local emotional changes but also of modeling global emotional evolution trends effectively.

From the perspective of emotional dependency evolution, emotions in long dialogues exhibit clear non-linear and phased characteristics. In the early stages, the model primarily relies on local modality features for judgment. As the dialogue progresses, emotional interactions between speakers gradually intensify, and the differential mechanism captures changes in the positive and negative subspaces. In the later stages, the model strengthens critical contextual information through dynamic attention, effectively integrating information across the entire sequence. This process ensures a comprehensive modeling of the emotional trajectory in the dialogue, further validating the robustness and adaptability of the proposed method in long-sequence dialogue tasks.

6.4. Computational overhead increment analysis

To comprehensively evaluate the computational impact of the proposed differential graph convolution network and the adaptive modality balancing module, we conduct a comparative analysis of the full model on the IEMOCAP and MELD datasets in terms of inference time, batch processing time, throughput, and floating-point operations per second (FLOPs), as reported in Table 8.

For the differential graph convolution network, compared with the conventional graph attention network, the additional computational overhead mainly arises from the introduction of the differential attention mechanism and relation embedding computation. By modeling the differences among multiple attention distributions, this mechanism effectively highlights context-relevant dependencies, thereby enhancing the modeling of intra-speaker emotional continuity and inter-speaker emotional interactions. As each node is required to compute multiple sets of attention weights and perform differential operations, the computational complexity increases with the graph size and the number of edges, resulting in a moderate increase in inference time and batch processing time, accompanied by a slight decrease in throughput. Notably, this overhead is primarily confined to attention computations within local subgraphs and does not introduce global high-order operations, leading to a relatively mild growth in complexity as the dialogue length increases. In long-sequence dialogue scenarios, DiffRGCN can still effectively limit the computational scope through window-based subgraph modeling, thereby maintaining good scalability.

For the adaptive modality balancing module, its core objective is to dynamically regulate the relative contributions of different modalities during multimodal fusion via an adaptive modality dropout strategy, alleviating the dominance of a single modality in the decision-making process. The module mainly consists of lightweight operations, including probability estimation, stochastic sampling, and feature scaling, without introducing additional complex network structures or large-scale matrix operations, and thus imposes a limited computational burden on the overall model. In long-sequence scenarios, its computational complexity scales approximately linearly with the sequence length, enabling the model to maintain stable computational efficiency as dialogue length increases. By dynamically adjusting modality contributions while preserv-

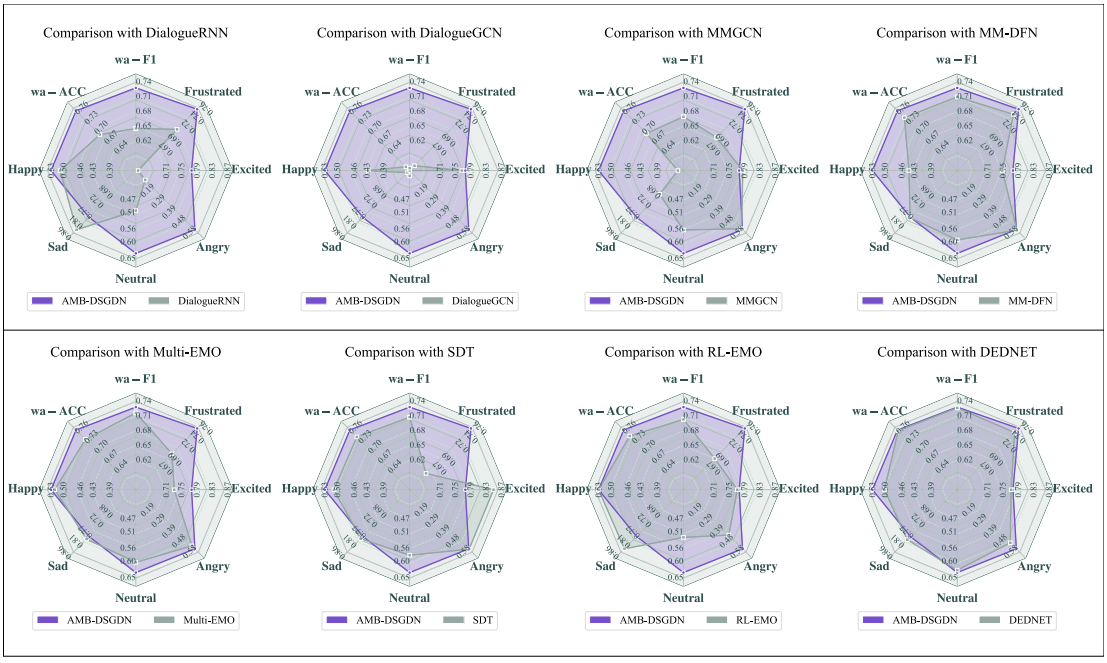

Fig. 6. Performance comparison of the proposed method with other baseline methods on the long sequence dialogue subset of the IEMOCAP dataset, intuitively showing the performance advantages of the proposed method in experiments.

Table 8
Computational overhead analysis of DiffRGCN compared to the standard GAT, and of the adaptive modality balancing module.

Module	Dataset	Inference Time (ms)		Throughput (samples/s)		Batch Time (ms)		FLOPs/s (T)	
		Baseline	Δ (Change)	Baseline	Δ (Change)	Baseline	Δ (Change)	Baseline	Δ (Change)
DiffRGCN	IEMOCAP	0.0769	+0.0090 (+11.70%)	13169.51	-1693.29 (-12.86%)	62.14	+10.7027 (+17.22%)	0.0805	-0.0157 (-19.50%)
	MELD	0.4271	+0.1026 (+24.03%)	2384.02	-404.02 (-16.95%)	60.33	+11.4914 (+19.05%)	0.0500	-0.00236 (-4.72%)
Adaptive Modality Balancing	IEMOCAP	0.0769	+0.0165 (+21.45%)	13169.51	-980.73 (-7.45%)	62.14	+6.52 (+10.50%)	0.0805	-0.0082 (-10.19%)
	MELD	0.4271	+0.0190 (+4.45%)	2384.02	-137.90 (-5.79%)	60.33	+2.1635 (+3.59%)	0.0500	-0.00384 (-7.68%)

ing computational efficiency, this module further enhances the model's generalization ability and robustness in complex conversational environments.

6.5. *Robustness analysis of multimodal noise:*

To evaluate the model's stability and generalization ability under complex noisy conditions, we conducted full-modality noise interference experiments on the IEMOCAP and MELD datasets. During testing, Gaussian noise of varying intensities was injected into the textual, visual, and audio features, with the amplitude normalized by the overall feature standard deviation to simulate perception errors, environmental disturbances, and modality quality fluctuations. The experiments covered multiple levels, from a noise-free baseline to high-intensity noise (0.1-0.7), and used weighted accuracy and weighted average F1-score as evaluation metrics.

As shown in Fig. 7, model performance declined smoothly rather than fluctuating sharply with increasing noise intensity, indicating strong robustness. On IEMOCAP, when full-modality noise reached 0.3, weighted accuracy and average F1-score only slightly decreased compared to the baseline, and even at 0.7 noise, the performance degradation remained controlled. On MELD, even under noise levels of 0.5-0.7, weighted accuracy and average F1-score remained around 65%, slightly lower than the baseline, demonstrating good noise resistance.

Further analysis indicates that DiffGCN consistently captures dynamic emotional dependencies both across modalities and within/between speakers. Its differential attention mechanism effectively reduces redundant inter-modality interference, preserving discriminative capa-

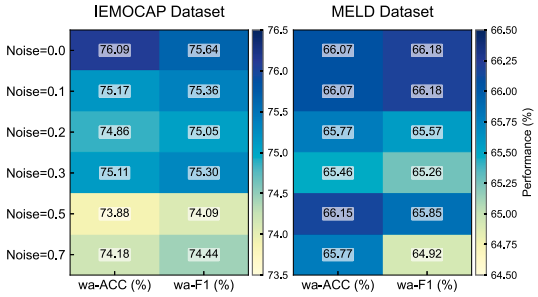

Fig. 7. Full-modality noise robustness test results of the model on the IEMOCAP and MELD datasets. Each cell in the figure shows the model's wa-ACC and wa-F1 under different noise intensities. The noise intensity ranges from 0 to 0.7, representing the standard deviation multiples of Gaussian noise added to the textual, visual, and audio features during testing. The color encoding indicates performance levels, with darker colors representing higher performance; the color bar on the right provides a reference for the corresponding percentage values. This heatmap intuitively illustrates the model's stability and performance variations under different noise conditions. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 9
Model parameters and runtime comparison.

Method	Params (M)	IEMOCAP Runtime (s)	MELD Runtime (s)
DialogueGCN	12.92	58.1	127.5
RGAT	15.28	68.5	146.3
LR-GCN	15.77	87.7	142.3
MMGCN	0.46	93.7	75.3
DER-GCN	78.59	125.5	189.7
AMB-DSGDN (ours)	13.13	179.95	146.4

bility under noisy conditions. Meanwhile, the adaptive modality balancing mechanism dynamically adjusts modality weights under high-noise conditions, preventing dominant modalities from overwhelming the fusion process while retaining useful information from minor modalities, thus maintaining consistency and reliability of the model outputs under full-modality noise.

6.6. Complexity analysis

Table 9 summarizes the experimental results of the compared models in terms of parameter count and runtime. From the experimental data, although our model has a parameter scale similar to some graph convolutional models, its inference time is slightly higher, which is directly related to the introduction of the differential relational graph convolutional network and the adaptive modality balancing module. Experimental analysis shows that the differential relational graph convolutional network performs multiple attention computations and integrates relational embeddings for each node, enhancing the ability to capture dependencies both within and across speakers, but it also brings additional computational overhead. The adaptive modality balancing module dynamically adjusts the contribution of each modality in each batch, ensuring that weak modality information is effectively utilized; experiments show that its additional impact on overall inference time is relatively limited. Although there is some increase in computational cost, compared to the performance improvement, this overhead remains within an acceptable range. The experimental results indicate that the model achieves a good balance between high accuracy, computational complexity, and multimodal information fusion.

6.7. Adaptive modality balancing under extreme imbalance

Based on the single-modality performance differences in Table 4, we conducted extreme modal imbalance experiments on the IEMOCAP and MELD datasets, setting the fusion weights for text, visual, and audio modalities to 0.8, 0.1, and 0.1, respectively. Results in Fig. 8 show that, despite the dominance of the text modality, the model maintains high overall accuracy and weighted F1 scores. Further analysis of the confusion matrices indicates that the visual and audio modalities still make significant contributions in recognizing key emotions such as “anger” and “excited,” effectively mitigating misclassifications caused by text modality bias. This demonstrates that the model can adaptively regulate the information flow based on the relative effectiveness of each modality, preventing the dominant modality from overwhelming the fusion result. Combined with the differential graph attention mechanism, this regulation also suppresses shared noise, ensuring a more balanced contribution from each modality at different stages. Overall, the experiments validate the effectiveness of the adaptive modality balancing mechanism in enhancing model robustness and sensitivity to dynamic multimodal emotional variations under extreme modal imbalance.

7. Conclusion

In this study, we propose AMB-DSGDN for multimodal conversation emotion recognition. The model effectively mitigates modality contribution imbalances through an adaptive modality dropout mechanism

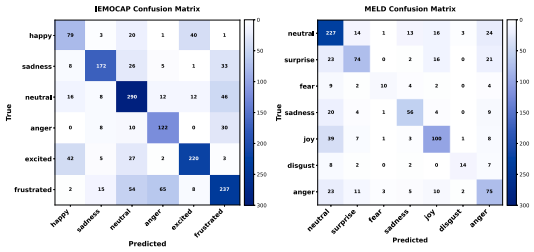

Fig. 8. This figure presents the emotion recognition confusion matrices for the IEMOCAP (left) and MELD (right) datasets under the extreme modality imbalance experiment: the left matrix corresponds to the 6 emotion categories of IEMOCAP, while the right one corresponds to the 7 emotion categories of MELD. The "True" axis represents the ground-truth emotion labels, and the "Predicted" axis represents the model’s predicted labels; the values inside the cells indicate the number of samples for the corresponding category. The dark-colored blocks along the matrix diagonal reflect the high recognition accuracy for most emotions, while the light-colored blocks off the diagonal reflect the misclassification between different emotions.

based on real-time performance evaluation. Meanwhile, it independently constructs inter-speaker and intra-speaker relational subgraphs for each modality and incorporates a DiffRGCN, which captures the dynamic evolution of emotional dependencies via positive-negative subspace projections and differential attention operations. Experimental results demonstrate that AMB-DSGDN significantly outperforms existing baseline methods on the IEMOCAP and MELD datasets, achieving stable improvements in overall performance and specific emotion categories.

However, due to the computational complexity and overhead introduced by graph-structured modeling, the inference efficiency of the proposed model remains limited when processing extremely long dialogue sequences and operating in resource-constrained environments. In particular, for deployment on edge devices, challenges such as model size, the computational cost of graph attention mechanisms, and the real-time processing of multimodal features need to be carefully addressed. Future work will focus on improving computational efficiency by exploring lightweight graph attention designs, subgraph pruning, and parameter sharing strategies, as well as investigating model compression, knowledge distillation, and hardware-aware acceleration techniques to enhance real-time inference performance on edge devices. These efforts are expected to further improve the applicability of the proposed framework in real-time human-machine interaction and emotion analysis scenarios.

CRediT authorship contribution statement

Yunsheng Wang: Conceptualization, Methodology, Data curation, Writing – original draft; **Yuntao Shou:** Supervision, Writing – review & editing; **Yilong Tan:** Supervision, Investigation, Writing – review & editing; **Wei Ai:** Supervision, Investigation, Writing – review & editing; **Tao Meng:** Supervision, Investigation, Writing – review & editing; **Keqin Li:** Supervision, Investigation, Writing – review & editing.

Data availability

Data will be made available on request.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

The authors deepest gratitude goes to the anonymous reviewers and AE for their careful work and thoughtful suggestions that have helped improve this paper substantially. This work is supported by National Natural Science Foundation of China (Grant No. 69189338), Excellent Young Scholars of Hunan Province of China (Grant No. 22B0275), and program of Research on Local Community Structure Detection Algorithms in Complex Networks (Grant No. 2020YJ009).

References

Ai, W., Shou, Y., Meng, T., & Li, K. (2024). Der-gcn: Dialog and event relation-aware graph convolutional neural network for multimodal dialog emotion recognition. *IEEE Transactions on Neural Networks and Learning Systems*, *36*(3), 4908–4921.

Ai, W., Zhang, F., Shou, Y., Meng, T., Chen, H., & Li, K. (2025). Revisiting multimodal emotion recognition in conversation from the perspective of graph spectrum. In *Proceedings of the AAAI conference on artificial intelligence* (pp. 11418–11426). (*vol. 39*).

Busso, C., Bulut, M., Lee, C.-C., Kazemzadeh, A., Mower, E., Kim, S., Chang, J. N., Lee, S., & Narayanan, S. S. (2008). Iemocap: Interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, *42*(4), 335–359.

Chandola, D., Altarawneh, E., Jenkin, M., & Papagelis, M. (2024). Serc-gcn: speech emotion recognition in conversation using graph convolutional networks. In *ICASSP 2024-2024 IEEE international conference on acoustics, speech and signal processing (icassp)* (pp. 76–80). IEEE.

Chen, J., Yang, D., Jiang, Y., Li, M., Wei, J., Hou, X., & Zhang, L. (2024). Efficiency in focus: Layernorm as a catalyst for fine-tuning medical visual language pre-trained models. arXiv preprint arXiv:2404.16385, .

Cheng, Z., Cheng, Z.-Q., He, J.-Y., Wang, K., Lin, Y., Lian, Z., Peng, X., & Hauptmann, A. (2024). Emotion-llama: Multimodal emotion recognition and reasoning with instruction tuning. *Advances in Neural Information Processing Systems*, *37*, 110805–110853.

Eyben, F., Wöllmer, M., & Schuller, B. (2010). Opensmile: The munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM international conference on multimedia* (pp. 1459–1462).

Fan, Q., Zuo, H., Liu, R., Lian, Z., & Gao, G. (2024). Learning noise-robust joint representation for multimodal emotion recognition under incomplete data scenarios. In *Proceedings of the 2nd international workshop on multimodal and responsible affective computing* (pp. 116–124).

Farhadipour, A., Ranjbar, H., Chapariniya, M., Vukovic, T., Ebling, S., & Dellwo, V. (2025). Multimodal emotion recognition and sentiment analysis in multi-party conversation contexts. arXiv preprint arXiv:2503.06805.

Fu, Y., Yan, X., Chen, W., & Zhang, J. (2025). Feature-enhanced multimodal interaction model for emotion recognition in conversation. *Knowledge-Based Systems*, *309*, 112876. https://doi.org/10.1016/j.knosys.2024.112876

Ghosal, D., Majumder, N., Poria, S., Chhaya, N., & Gelbukh, A. (2019). Dialoguegcn: A graph convolutional neural network for emotion recognition in conversation. arXiv preprint arXiv:1908.11540.

Guo, L., Song, Y., & Ding, S. (2024a). Speaker-aware cognitive network with cross-modal attention for multimodal emotion recognition in conversation. *Knowledge-Based Systems*, *296*, 111969.

Guo, Z., Jin, T., & Zhao, Z. (2024b). Multimodal prompt learning with missing modalities for sentiment analysis and emotion recognition. arXiv preprint arXiv:2407.05374.

Hu, D., Hou, X., Wei, L., Jiang, L., & Mo, Y. (2022). Mm-dfn: multimodal dynamic fusion network for emotion recognition in conversations. In *ICASSP 2022-2022 IEEE international conference on acoustics, speech and signal processing (icassp)* (pp. 7037–7041). IEEE.

Hu, J., Liu, Y., Zhao, J., & Jin, Q. (2021). Mmgcn: Multimodal fusion via deep graph convolution network for emotion recognition in conversation. arXiv preprint arXiv:2107.06779.

Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2017). Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4700–4708).

Joshi, A., Bhat, A., Jain, A., Singh, A. V., & Modi, A. (2022). Cogmen: Contextualized gnn based multimodal emotion recognition. arXiv preprint arXiv:2205.02455.

Kong, H., Lou, X., & Li, Z. (2024). Emotional EEG recognition and classification based on GCN combined with LSTM. In *2024 IEEE international conference on medical artificial intelligence (medAI)* (pp. 294–301). IEEE.

Li, J., Wang, X., Lv, G., & Zeng, Z. (2023). GraphCFC: A directed graph based cross-modal feature complementation approach for multimodal conversational emotion recognition. *IEEE Transactions on Multimedia*, *26*, 77–89.

Lian, H., Lu, C., Li, S., Zhao, Y., Tang, C., & Zong, Y. (2023). A survey of deep learning-based multimodal emotion recognition: speech, text, and face. *Entropy*, *25*(10), 1440.

Lian, Z., Chen, H., Chen, L., Sun, H., Sun, L., Ren, Y., Cheng, Z., Liu, B., Liu, R., Peng, X. et al. (2025). Affectgpt: A new dataset, model, and benchmark for emotion understanding with multimodal large language models. arXiv preprint arXiv:2501.16566.

Lian, Z., Sun, H., Sun, L., Wen, Z., Zhang, S., Chen, S., Gu, H., Zhao, J., Ma, Z., Chen, X. et al. (2024a). Mer 2024: Semi-supervised learning, noise robustness, and open-vocabulary multimodal emotion recognition. In *Proceedings of the 2nd international workshop on multimodal and responsible affective computing* (pp. 41–48).

Lian, Z., Sun, L., Ren, Y., Gu, H., Sun, H., Chen, L., Liu, B., & Tao, J. (2024b). Merbench: A unified evaluation benchmark for multimodal emotion recognition. arXiv preprint arXiv:2401.03429.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692.

Liu, Y., Zhang, H., Zhan, Y., Chen, Z., Yin, G., Wei, L., & Chen, Z. (2024). Noise-resistant multimodal transformer for emotion recognition. *International Journal of Computer Vision*, *133* (pp. 1–21).

Ma, H., Wang, J., Lin, H., Zhang, B., Zhang, Y., & Xu, B. (2024). A transformer-based model with self-distillation for multimodal emotion recognition in conversations. *IEEE Transactions on Multimedia*, *26*, 776–788. https://doi.org/10.1109/TMM.2023.3271019

Majumder, N., Poria, S., Hazarika, D., Mihalcea, R., Gelbukh, A., & Cambria, E. (2019). Dialoguernn: An attentive rnn for emotion detection in conversations. In *Proceedings of the AAAI conference on artificial intelligence* (pp. 6818–6825). (*vol. 33*).

Mehrez, H., & Selouani, S. A. (2025). Multimodal emotion recognition for conversational systems in continuous affective space. In *Proceedings of the 6th international conference on bio-engineering for smart technologies (bioSMART)* (pp. 1–4).

Poria, S., Hazarika, D., Majumder, N., Naik, G., Cambria, E., & Mihalcea, R. (2018). Meld: A multimodal multi-party dataset for emotion recognition in conversations. arXiv preprint arXiv:1810.02508.

Reddy, G. R. K., Bhavani, A. D., & Odugu, V. K. (2025). Optimized recurrent neural network based brain emotion recognition technique. *Multimedia Tools and Applications*, *84*(8), 4655–4674.

Shi, T., & Huang, S.-L. (2023). MultiEMO: An attention-based correlation-aware multimodal fusion framework for emotion recognition in conversations. In *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 14752–14766).

Shou, Y., Meng, T., Ai, W., & Li, K. (2024). Dynamic graph neural ode network for multimodal emotion recognition in conversation. arXiv preprint arXiv:2412.02935.

Sun, T., Wei, Y., Ni, J., Liu, Z., Song, X., Wang, Y., & Nie, L. (2024). Muti-modal emotion recognition via hierarchical knowledge distillation. *IEEE Transactions on Multimedia*, *26*, 9036–9046.

Sun, Y., & Zhou, T. (2025). DialogueMLLM: Transforming multimodal emotion recognition in conversation through instruction-tuned MLLM. *IEEE Access*, *13*, 1–12. https://doi.org/10.1109/ACCESS.2025.XXXXX

Tu, G., Xiong, F., Liang, B., Wang, H., Zeng, X., & Xu, R. (2024). Multimodal emotion recognition calibration in conversations. In *Proceedings of the 32nd ACM international conference on multimedia* (pp. 9621–9630).

Wang, H., Yang, W., Zhong, X., Zhou, J., Liu, F., & Zhang, W. (2025). Mitigating modality quantity and quality imbalance in multimodal online federated learning. arXiv preprint arXiv:2508.11159.

Wang, Y., Zhang, W., Liu, K., Wu, W., Hu, F., Yu, H., & Wang, G. (2024). Dynamic emotion-dependent network with relational subgraph interaction for multimodal emotion recognition. *IEEE Transactions on Affective Computing*, *16*, 712–725. .

Wei, Y., Feng, R., Wang, Z., & Hu, D. (2024). Enhancing multimodal cooperation via sample-level modality valuation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 27338–27347).

Wu, C., Cai, Y., Liu, Y., Zhu, P., Xue, Y., Gong, Z., Hirschberg, J., & Ma, B. (2025a). Multimodal emotion recognition in conversations: A survey of methods, trends, challenges and prospects. arXiv preprint arXiv:2505.20511.

Wu, Y., Zhang, S., & Li, P. (2025b). Multi-modal emotion recognition in conversation based on prompt learning with text-audio fusion features. *Scientific Reports*, *15*(1), 8855. https://doi.org/10.1038/s41598-025-08855-4

Zhang, C., Zhang, Y., & Cheng, B. (2024). Rl-emo: A reinforcement learning framework for multimodal emotion recognition. In *ICASSP 2024-2024 IEEE international conference on acoustics, speech and signal processing (icassp)* (pp. 10246–10250). IEEE.

Zhao, H., Gao, Y., Chen, H., Li, B., Ye, G., & Zhang, Z. (2025a). Enhanced multimodal emotion recognition in conversations via contextual filtering and multi-frequency graph propagation. In *ICASSP 2025-2025 IEEE international conference on acoustics, speech and signal processing (icassp)* (pp. 1–5). IEEE.

Zhao, S., Ren, J., & Zhou, X. (2025b). Cross-modal gated feature enhancement for multimodal emotion recognition in conversations. *Scientific Reports*, *15*(1), 30004. https://doi.org/10.1038/s41598-025-30004-0