G2SERC: Unifying Graph Context Encoding and Sequential Emotion Decoding for Emotion Recognition in Conversation

Anonymous ACL submission

Abstract

Emotion Recognition in Conversation (ERC) aims to infer the emotions expressed in dialogues, where a relatively stable dialogue-level affective context coexists with transient utterance-level dynamics. Prior work often emphasizes either sequential modeling of local context or graph-based aggregation of global dependencies, leaving the interaction between global atmosphere and evolving utterance emotions under-explored. In this paper, we propose G2SERC, a graph-to-sequence framework that unifies relation-aware graph context encoding and sequential emotion decoding. G2SERC first builds a speaker-aware heterogeneous graph over conversation and employs a relation-aware graph encoder to derive dialogue-level affective context and speaker-level affective priors. A coupled recurrent decoder then tracks utterance dynamics while updating speaker-specific affective states, enabling emotion prediction conditioned on both dialogue evolution and speaker trajectories. Extensive experiments show that G2SERC consistently outperforms strong baselines and achieves state-of-the-art performance. Additional analyses demonstrate improved robustness to local emotional perturbations and substantiate the benefit of integrating global and speaker-aware signals. [1]

1 Introduction

Emotion Recognition in Conversation (ERC) (Poria et al., 2019b; Fu et al., 2023) aims to identify the emotional states expressed throughout a dialogue, and it is central to applications such as dialogue systems (Jiao et al., 2020b; Gong et al., 2023), social media analysis (Kumar et al., 2015), and human-computer interaction (Hu et al., 2021b). Emotions in conversations exhibit structured variations across contextual levels (Ai et al., 2025): a dialogue often maintains a relatively stable, dialogue-level affective atmosphere that reflects the overall emotional tone of the interaction, while individual utterances express transient emotions that evolve in response to local contextual triggers. As illustrated in Figure 1, a conversation may preserve a positive overall atmosphere even when individual utterances temporarily deviate due to contextual perturbations. These observations suggest that utterance-level emotion recognition can benefit from incorporating dialogue-level affective context that captures the overall atmosphere.

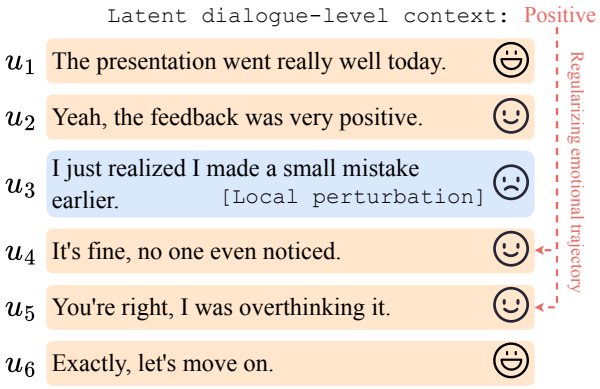

Figure 1: An illustrative example showing that dialogue-level affective context supports stable emotion prediction under local emotional perturbations.

To capture the interplay between dialogue-level affective context and utterance-level emotional dynamics, existing ERC methods have adopted diverse modeling paradigms (Poria et al., 2019a; Wu et al., 2025). Implicit dependency modeling methods, including sequence-based (Poria et al., 2017; Hazarika et al., 2018; Majumder et al., 2019; Hu et al., 2021b, 2022b; Gou et al., 2025) and Transformer-based (Zhong et al., 2019; Li et al., 2020; Shen et al., 2021a; Mao et al., 2021) approaches, encode conversational dependencies by folding dialogue-level information into evolving hidden representations via recurrent transitions or self-attention. While effective for modeling local emotional continuity, these implicit representations

[1] The code is released at https://anonymous.4open.science/r/cjkdgs29.

are sensitive to local perturbations and lack a stable dialogue-level affective anchor to guide subsequent emotional evolution. In contrast, explicit dependency modeling methods, most notably graph-based ERC approaches (Ghosal et al., 2019; Shen et al., 2021b; Chen et al., 2023; Tu et al., 2024; Ai et al., 2025; Li et al., 2025), define utterance interactions through structured relational edges and aggregate conversational context via message passing. However, the aggregated information is typically compressed into utterance-level representations for point-wise emotion prediction, without explicitly modeling how dialogue-level affective context shapes emotion trajectories over time. Consequently, existing methods tend to emphasize either implicit temporal dynamics or explicit contextual aggregation, but rarely offer a unified mechanism that leverages both in a coordinated manner.

In this work, we aim to jointly leverage dialogue-level affective context and utterance-level emotional dynamics within a unified framework. Our motivation is that graph-based models excel at aggregating global conversational context, whereas sequence-based models are effective at capturing utterance-level emotional evolution. Building on this complementarity, we present G2SERC, a graph-to-sequence framework in which graph-encoded global context guides sequential emotion modeling. Specifically, G2SERC comprises a graph-based context encoder and a sequential emotion decoder: the encoder aggregates heterogeneous conversational dependencies to model dialogue-level affective context, and the decoder models utterance-level emotional dynamics conditioned on the encoded global context. By decoupling contextual aggregation from temporal emotion modeling, our framework enables coordinated use of global affective context and local emotional dynamics.

We conduct extensive experiments on four public ERC benchmarks, and the results consistently demonstrate the effectiveness of G2SERC compared to several baselines. Moreover, an emotion transition analysis provides additional evidence that G2SERC maintains stable emotion prediction for subsequent utterances despite transient local emotional deviations.

The contributions of this work are: **(1)** We propose a graph-to-sequence framework that jointly models dialogue-level affective context and utterance-level emotional dynamics for ERC. **(2)** We design a graph-based encoder to capture global affective information from conversational structures. **(3)** We develop a coupled sequential decoder that leverages graph-encoded global context to guide utterance-level emotion prediction. **(4)** Extensive experiments on multiple benchmark datasets further validate the effectiveness of the proposed approach.

2 Related Work

2.1 Context Modeling in Conversation

Conversations are characterized by complex relational structures arising from multi-party interactions (Fu et al., 2023; Wu et al., 2025), which distinguishes ERC from conventional sentiment analysis (Yin and Zhong, 2024; Miah et al., 2024; Li et al., 2024b; Sharma et al., 2025). Early ERC studies (Poria et al., 2017; Majumder et al., 2019; Jiao et al., 2020a; Zhao et al., 2022) primarily modeled contextual dependencies across dialogue turns with recurrent architectures, capturing emotion transitions along the temporal dimension (Hazarika et al., 2018; Ghosal et al., 2020; Hu et al., 2021a). Despite their effectiveness on short-range contextual cues, sequence-based approaches often struggle to represent non-linear conversational structures, including long-range or cross-speaker dependencies (Ghosal et al., 2019). To improve long-distance contextual reasoning, subsequent works adopted Transformer architectures (Wang et al., 2020; Zhu et al., 2021; Mao et al., 2021; Khule et al., 2024) and pretrained language models (Yu et al., 2024; Lei et al., 2024; Jing et al., 2026), where self-attention implicitly aggregates conversational context (Zhong et al., 2019; Shen et al., 2021a; Wu et al., 2024). More recently, graph-based methods (Ghosal et al., 2019; Fu et al., 2021, 2022; Li et al., 2023; Chen et al., 2023; Ai et al., 2025; Shou et al., 2025; Li et al., 2025) have explicitly modeled conversational context by constructing graphs over utterances, with relational edges encoding structural dependencies. Such graph representations naturally capture conversational topology and facilitate the aggregation of global contextual information, enabling a more holistic understanding of the dialogue (Zhang et al., 2019; Hu et al., 2021b; Tu et al., 2024). Extensions that incorporate external knowledge further enrich these representations (Zhong et al., 2019; Ghosal et al., 2020; Fu et al., 2021).

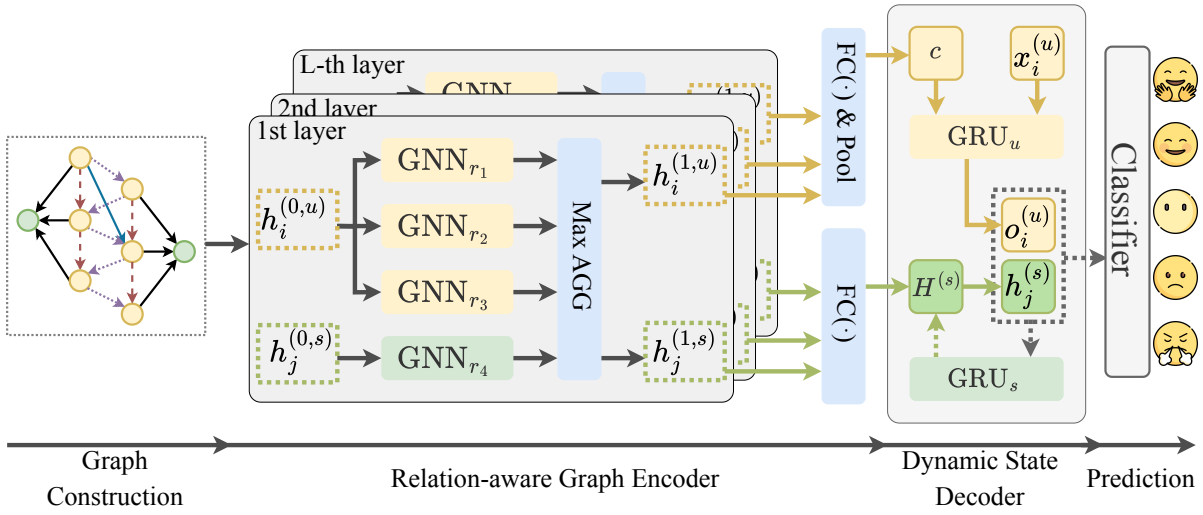

Figure 2: Overall architecture of our G2SERC.

2.2 Emotional Dynamics Modeling

A substantial line of research focuses on modeling emotional dynamics at the utterance level, where intra-speaker emotional continuity and inter-speaker emotional influence jointly shape emotion evolution (Poria et al., 2019b; Ghosal et al., 2019). Early approaches leveraged recurrent or attention-based architectures (Hazarika et al., 2018; Majumder et al., 2019; Wang et al., 2020; Li et al., 2020; Ghosal et al., 2020; Hu et al., 2021a; Shen et al., 2021a; Hu et al., 2023) to track emotion trajectories over time. Subsequent studies enhanced utterance-level dynamics modeling by incorporating inter-speaker interaction patterns (Bao et al., 2022), pretrained contextual representations (Khule et al., 2024; Lei et al., 2024), and personality-related attributes (Wang et al., 2024). Graph-based models have introduced speaker-related signals largely through static structural designs, for example by augmenting utterance representations with speaker embeddings (Hu et al., 2021b, 2022a; Li et al., 2023; Chen et al., 2023; Tu et al., 2024; Ai et al., 2025; Shou et al., 2025) or by explicitly modeling speakers as nodes connected to utterances (Zhang et al., 2019; Song et al., 2023). Relational edges have also been used to encode intra- and inter-speaker interactions (Ghosal et al., 2019; Fu et al., 2021; Shen et al., 2021b). In contrast, our G2SERC adopts a unified perspective that explicitly integrates dialogue-level affective context with utterance-level emotional dynamics, providing a principled way to relate global conversational affect to local emotional evolution within a single framework.

3 Methodology

In this section, we introduce G2SERC for ERC. As shown in Figure 2, G2SERC comprises four components: (1) a heterogeneous *Graph Construction* module (§3.2); (2) a *Relation-aware Graph Encoder* for contextual representation learning (§3.3); (3) a *Dynamic State Decoder* that models utterance- and speaker-specific temporal dynamics (§3.4); and (4) a *Prediction* head (§3.5).

3.1 Preliminary

Task Definition. Given a conversation represented as an ordered sequence of N utterances $\mathcal{U} = \{u_1, u_2, \ldots, u_N\}$, each utterance u_i is produced by a speaker $s_{\pi(u_i)}$, where $\pi(\cdot)$ maps an utterance to the index of its corresponding speaker. The conversation involves M distinct speakers ($M \geq 2$), denoted as $\{s_1, \ldots, s_M\}$. The goal of ERC is to assign an emotion label $y_i \in \mathcal{C}$ to each utterance u_i, where $\mathcal{C}$ is the set of predefined emotion categories.

Textual Feature Extraction. We encode each utterance $u_i \in \mathcal{U}$ into a contextual representation $\boldsymbol{x}_i^{(u)} \in \mathbb{R}^{d_h}$ using a pretrained RoBERTa (Liu et al., 2019) model, where d_h denotes the dimensionality of the utterance representation. The utterance encoder is kept fixed during training to stabilize optimization and to decouple utterance-level representation learning from conversation-level modeling in our framework.

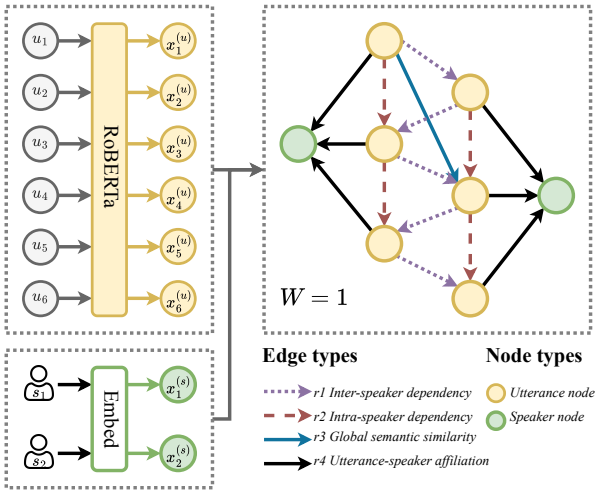

Figure 3: An illustrative example of the heterogeneous conversation graph with window size $W = 1$, consisting of utterance and speaker nodes connected by four types of relations ($r1$-$r4$).

3.2 Graph Construction

To explicitly model heterogeneous conversational dependencies across utterances and speakers, we represent each conversation as a speaker-aware heterogeneous graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{R})$, as shown in Figure 3. The node set $\mathcal{V}$ consists of two disjoint subsets, i.e., $\mathcal{V} = \mathcal{V}^{(u)} \cup \mathcal{V}^{(s)}$, where $\mathcal{V}^{(u)}$ and $\mathcal{V}^{(s)}$ denote utterance nodes and speaker nodes, respectively. Edges $(v_i, v_j) \in \mathcal{E}$ encode structured conversational dependencies, and each edge is associated with a relation type $r \in \mathcal{R}$.

Nodes. Following prior work (Song et al., 2023), we include two types of nodes: utterance nodes and speaker nodes. Each utterance $u_i \in \mathcal{U}$ corresponds to an utterance node $v_i^{(u)}$, initialized with its RoBERTa representation $\boldsymbol{x}_i^{(u)}$. Each speaker s_j corresponds to a speaker node $v_j^{(s)}$, initialized with a learnable embedding $\boldsymbol{x}_j^{(s)} \in \mathbb{R}^{d_h}$ that encodes speaker identity information.

Edges. Conversational dependencies are heterogeneous in nature (Ghosal et al., 2019; Shen et al., 2021b), motivating multiple relation types $r \in \mathcal{R}$: ($r1$) *Inter-speaker dependency* captures short-range cross-speaker influence. We add a directed edge $(v_i^{(u)}, v_j^{(u)})$ if $1 \le j - i \le W$ and $\pi(u_i) \neq \pi(u_j)$. ($r2$) *Intra-speaker dependency* models speaker-specific emotional continuity independent of turn-taking. we add a directed edge $(v_i^{(u)}, v_j^{(u)})$ if u_i and u_j are consecutive utterances produced by the same speaker.

($r3$) *Global semantic similarity* enables sparse long-range information exchange beyond local context. We add an undirected edge $(v_i^{(u)}, v_j^{(u)})$ if $i \neq j$ and $\cos\left(\boldsymbol{x}_i^{(u)}, \boldsymbol{x}_j^{(u)}\right) > \tau_s$, where τ_s is a predefined threshold.

($r4$) *Utterance-speaker affiliation* facilitates explicit interaction between utterance representations and speaker states. For each utterance u_i, we add a directed edge $(v_i^{(u)}, v_{\pi(u_i)}^{(s)})$.

3.3 Relation-aware Graph Encoder

The relation-aware graph encoder aggregates structured conversational context into dialogue-level representations, without modeling utterance-level temporal dynamics. In the conversation graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{R})$, different relation types r encode heterogeneous contextual signals that may be obscured if all edges are mixed within a single message-passing process. Accordingly, we perform message propagation separately for each relation type and then integrate their contributions in a controlled manner.

Relation-aware Context Propagation. Given the relation set $\mathcal{R}$, we conduct message propagation over edges associated with each relation $r \in \mathcal{R}$ independently. At the l-th layer, for a target node $v_i \in \mathcal{V}$, let $\mathcal{N}_r(v_i)$ denote the set of neighbors connected to v_i via relation r. We update the relation-specific representation of v_i as:

$$\boldsymbol{h}_i^{(l+1,r)} = \mathrm{GNN}_r^{(l)}\left(\boldsymbol{h}_i^{(l)}, \sum_{v_j \in \mathcal{N}_r(v_i)} \boldsymbol{h}_j^{(l)}\right), \quad (1)$$

where $\mathrm{GNN}_r^{(l)}(\cdot)$ is a relation-specific graph neural network that aggregates messages from neighbors under relation r. Here, $\boldsymbol{h}_i^{(l)}$ denotes the representation of node v_i at layer l, and $\boldsymbol{h}_i^{(0)} = \boldsymbol{x}_i$ is the input node feature. In this work, we instantiate $\mathrm{GNN}_r^{(l)}(\cdot)$ using graph attention networks (GAT) (Brody et al., 2022), where attention coefficients modulate the contribution of each neighbor under relation r. For an edge $(v_j, v_i)_r$, the attention coefficient $\alpha_{ij}^{(r)}$ is computed as:

$$\begin{aligned} \gamma_{ij}^{(r)} &= \boldsymbol{a}_r^\top \sigma\left(\boldsymbol{W}_r \left[\boldsymbol{h}_i^{(l)}, \boldsymbol{h}_j^{(l)}\right]\right), \\ \alpha_{ij}^{(r)} &= \mathrm{softmax}_{j \in \mathcal{N}_r(v_i)}\left(\gamma_{ij}^{(r)}\right), \end{aligned} \quad (2)$$

where $\boldsymbol{W}_r \in \mathbb{R}^{d_h \times 2d_h}$ and $\boldsymbol{a}_r \in \mathbb{R}^{d_h}$ are learnable parameters associated with relation r, and $\sigma(\cdot)$ denotes the LeakyReLU activation function.

Multi-relation Context Fusion. After relation-aware propagation, each node v_i obtains a set of relation-specific representations $\{\boldsymbol{h}_i^{(l+1,r)} \mid r \in \mathcal{R}\}$. corresponding to different dependency relations in $\mathcal{G}$. To integrate these heterogeneous signals while avoiding interference across relations, we apply an element-wise max operator:

$$\widetilde{\boldsymbol{h}}_i^{(l+1)} = \mathrm{AGG}\big(\{\boldsymbol{h}_i^{(l+1,r)} \mid r \in \mathcal{R}\}\big), \quad (3)$$

where $\mathrm{AGG}(\cdot)$ denotes max aggregation. We then apply a non-linear activation followed by layer normalization to obtain the final node representation $\boldsymbol{h}_i^{(l+1)} \in \mathbb{R}^{d_h}$:

$$\boldsymbol{h}_i^{(l+1)} = \mathrm{LayerNorm}\big(\mathrm{ReLU}(\widetilde{\boldsymbol{h}}_i^{(l+1)})\big). \quad (4)$$

Notably, utterance nodes aggregate messages propagated through $r1$-$r3$, whereas speaker nodes are updated exclusively via the utterance–speaker relation $r4$, ensuring that speaker representations are informed only through their affiliated utterances.

Multi-layer Aggregation. We stack the relation-aware propagation and fusion operations for L layers to obtain progressively refined node representations. Since different graph layers capture context at different receptive fields, we aggregate representations across all layers. At layer l, we denote the representations of utterance nodes and speaker nodes as $\boldsymbol{H}^{(l,u)} = [\boldsymbol{h}_1^{(l,u)}; \cdots; \boldsymbol{h}_N^{(l,u)}] \in \mathbb{R}^{N \times d_h}$ and $\boldsymbol{H}^{(l,s)} = [\boldsymbol{h}_1^{(l,s)}; \cdots; \boldsymbol{h}_M^{(l,s)}] \in \mathbb{R}^{M \times d_h}$, respectively. We concatenate the layer-wise representations as:

$$\begin{aligned} \widetilde{\boldsymbol{H}}^{(u)} &= \Big[\boldsymbol{H}^{(0,u)}; \cdots; \boldsymbol{H}^{(L,u)}\Big], \\ \widetilde{\boldsymbol{H}}^{(s)} &= \Big[\boldsymbol{H}^{(0,s)}; \cdots; \boldsymbol{H}^{(L,s)}\Big]. \end{aligned} \quad (5)$$

We then apply linear projections to obtain compact utterance-level and speaker-level representations:

$$\begin{aligned} \boldsymbol{H}^{(u)} &= \mathrm{FC}_u\big(\widetilde{\boldsymbol{H}}^{(u)}\big), \\ \boldsymbol{H}^{(s)} &= \mathrm{FC}_s\big(\widetilde{\boldsymbol{H}}^{(s)}\big), \end{aligned} \quad (6)$$

where $\mathrm{FC}_u(\cdot)$ and $\mathrm{FC}_s(\cdot)$ are linear transformations with learnable parameters. Each row $\boldsymbol{h}_j^{(s)}$ of $\boldsymbol{H}^{(s)}$ serves as the encoded affective state of speaker s_j. To derive a dialogue-level representation, we apply max pooling over $\boldsymbol{H}^{(u)}$:

$$\boldsymbol{c} = \mathrm{Pool}(\boldsymbol{H}^{(u)}), \quad (7)$$

where $\mathrm{Pool}(\cdot)$ denotes max pooling. The resulting vector $\boldsymbol{c} \in \mathbb{R}^{d_h}$ represents a dialogue-level affective context that is permutation-invariant with respect to utterance order, providing a stable contextual prior for subsequent temporal decoding.

3.4 Dynamic State Decoder

Given the dialogue-level affective context $\boldsymbol{c}$ and the encoder-derived speaker states $\boldsymbol{H}^{(s)}$, the decoder models utterance-level temporal dynamics conditioned on global context, while tracking the evolution of speaker-specific affective states as the dialogue unfolds. We assume that utterance representations evolve sequentially over turns, and that a speaker's affective state is updated only when the speaker produces an utterance. Based on these assumptions, we adopt a dynamic decoding framework that jointly maintains utterance-level and speaker-level states.

Utterance-level Temporal Dynamics. We model utterance-level temporal dynamics by maintaining a latent dialogue state $\boldsymbol{o}_i^{(u)}$ for each utterance u_i, which captures sequential dependencies across dialogue turns. At each turn, the decoder updates the dialogue state based on the utterance representation $\boldsymbol{x}_i^{(u)}$. To capture bidirectional temporal dependencies, we use a bidirectional gated recurrent unit (BiGRU):

$$\begin{aligned} \boldsymbol{o}_i^{(u,f)} &= \mathrm{GRU}_u^f\big(\boldsymbol{x}_i^{(u)}, \boldsymbol{o}_{i-1}^{(u,f)}\big), \\ \boldsymbol{o}_i^{(u,b)} &= \mathrm{GRU}_u^b\big(\boldsymbol{x}_i^{(u)}, \boldsymbol{o}_{i+1}^{(u,b)}\big), \\ \boldsymbol{o}_i^{(u)} &= \big[\boldsymbol{o}_i^{(u,f)}; \boldsymbol{o}_i^{(u,b)}\big]. \end{aligned} \quad (8)$$

We initialize the forward and backward hidden states with the dialogue-level affective prior, i.e., $\boldsymbol{o}_0^{(u,f)} = \boldsymbol{o}_{N+1}^{(u,b)} = \boldsymbol{c}$, thereby conditioning temporal decoding on the global affective context $\boldsymbol{c}$.

Speaker-specific State Dynamics. We further assume that the affective state of a speaker $s_{\pi(u_i)}$ is updated only after the speaker produces utterance u_i. Accordingly, we treat the encoder-derived speaker representations $\boldsymbol{H}^{(s)}$ as initial speaker states and update them online during decoding. At turn i, given the utterance-level state $\boldsymbol{o}_i^{(u)}$ and its speaker index $j = \pi(u_i)$, we retrieve the current speaker state $\boldsymbol{h}_j^{(s)} \in \boldsymbol{H}^{(s)}$ and update it as:

$$\boldsymbol{h}_j^{(s)} \leftarrow \mathrm{GRU}_s\big(\boldsymbol{o}_i^{(u)}, \boldsymbol{h}_j^{(s)}\big), \quad (9)$$

while keeping the states of all other speakers unchanged.

Statistic	IEMOCAP	MELD	EmoryNLP	DailyDialog
Conversations (train)	120	1039	713	11118
Conversations (val)	–	114	99	1000
Conversations (test)	31	280	85	1000
Utterances (train)	5810	9989	9934	87170
Utterances (val)	–	1109	1344	8069
Utterances (test)	1623	2610	1328	7740
Average Turns	49.2	9.6	11.5	7.9
Average Speakers	2	2.7	3.2	2

Table 1: Dataset statistics.

3.5 Classification

For each utterance u_i, we predict its emotion label by jointly considering the utterance-level dynamic state $\boldsymbol{o}_i^{(u)}$ and the corresponding speaker state $\boldsymbol{h}_{\pi(u_i)}^{(s)}$ maintained *before* observing u_i. Formally, we concatenate the two states and feed them into a linear classifier:

$$\begin{aligned} \boldsymbol{o}_i &= \left[\boldsymbol{o}_i^{(u)}; \boldsymbol{h}_{\pi(u_i)}^{(s)}\right], \\ \boldsymbol{p}_i &= \mathrm{softmax}\left(\boldsymbol{W}_c \boldsymbol{o}_i + \boldsymbol{b}_c\right), \\ \widehat{y}_i &= \arg\max_{c \in \mathcal{C}} \boldsymbol{p}_i[c]. \end{aligned} \tag{10}$$

where $\boldsymbol{p}_i \in \mathbb{R}^{|\mathcal{C}|}$ denotes the predicted posterior distribution over emotion categories, and $\boldsymbol{W}_c \in \mathbb{R}^{|\mathcal{C}| \times 3d_h}$ and $\boldsymbol{b}_c \in \mathbb{R}^{|\mathcal{C}|}$ are trainable classifier parameters.

We train G2SERC with the standard cross-entropy loss:

$$\mathcal{L} = -\sum_{d=1}^{|\mathcal{D}|} \sum_{i=1}^{N_d} \log \boldsymbol{p}_i^{(d)} \left[y_i^{(d)}\right], \tag{11}$$

where $\mathcal{D}$ denotes the dialogue dataset, N_d is the number of utterances in the d-th conversation, and $y_i^{(d)}$ is the corresponding ground truth label.

4 Experimental Setups

4.1 Datasets

We evaluate our model on four widely used benchmark datasets for emotion recognition in conversation: **IEMOCAP** (Busso et al., 2008), **MELD** (Poria et al., 2019a), **EmoryNLP** (Zahiri and Choi, 2017), and **DailyDialog** (Li et al., 2017). Dataset statistics are summarized in Table 1, and detailed dataset descriptions and evaluation protocols are provided in Appendix A.

To ensure a consistent experimental setting across datasets, we use only the textual modality for all experiments. Following prior work, we report micro-averaged F1-score (micro-F1) on DailyDialog and weighted-average F1-score (w-F1) on the remaining datasets.

4.2 Baselines

We compare G2SERC with representative approaches from four categories: *(1) Sequence-based methods*, including DialogueRNN (Majumder et al., 2019), SGED (Bao et al., 2022), and SACL-LSTM (Hu et al., 2023); *(2) Transformer-based methods*, including DialogXL (Shen et al., 2021a), MultiEMO (Shi and Huang, 2023), and CFN-ESA (Li et al., 2024a); *(3) Graph-based methods*, including DialogueGCN (Ghosal et al., 2019), DAG-ERC (Shen et al., 2021b), and GS-MCC (Ai et al., 2025); and *(4) PLM-based methods*, including EmoBERTa (Kim and Vossen, 2021), ERC-DP (Wang et al., 2024), and AffectGPT-R1 (Lian et al., 2025). Detailed descriptions of these baselines are provided in Appendix B.

4.3 Training Details

All experiments are implemented in PyTorch and optimized with the AdamW. Detailed hyperparameter settings are provided in Appendix C. All experiments are conducted on a machine equipped with an NVIDIA GeForce RTX 5090 GPU, using CUDA 12.8 for acceleration.

5 Results and Analysis

5.1 Overall Results

Table 2 summarizes the overall performance of our proposed G2SERC and the compared baselines. Across all four benchmarks, G2SERC achieves the best results and establishes a new state-of-the-art on IEMOCAP, MELD, EmoryNLP, and DailyDialogue. These results demonstrate the effectiveness of our graph-to-sequence framework for ERC.

Compared with *sequence-based* and *transformer-based* methods, G2SERC yields clear and consistent gains. Notably, sequence-based models that incorporate additional signals (e.g., SGED and SACL+LSTM) also improve over their vanilla counterparts, highlighting the benefit of leveraging richer contextual information beyond pure sequential modeling. *Graph-based* approaches are generally stronger than purely sequential or transformer-based methods, suggesting that explicitly modeling conversational structures is beneficial. Building on these strengths, G2SERC further improves over graph-based baselines by

Method	IEMOCAP	MELD	EmoryNLP	DailyDialogue
Sequence-based Methods				
DialogueRNN (Majumder et al., 2019)	62.75	-	-	-
+RoBERTa	64.76	63.61	37.44	57.32
SGED (Bao et al., 2022)	68.53	65.46	40.24	-
SACL+LSTM (Hu et al., 2023)	69.22	66.45	39.65	-
Transformer-based Methods				
DialogXL (Shen et al., 2021a)	65.94	62.41	34.73	54.93
MultiEMO (Shi and Huang, 2023)	64.48	61.23	-	-
CFN-ESA (Li et al., 2024a)	66.57	65.81	-	-
Graph-based Methods				
DialogueGCN (Ghosal et al., 2019)	64.18	58.10	-	-
+RoBERTa	64.91	63.02	38.10	57.52
DAG-ERC (Shen et al., 2021b)	68.03	63.65	39.02	59.33
GS-MCC (Ai et al., 2025)	66.00	62.50	-	-
PLM-based Methods				
EmoBERTa (Kim and Vossen, 2021)	68.57	66.51	-	-
ERC-DP (Wang et al., 2024)	69.64	67.34	40.10	-
AffectGPT-R1 (Lian et al., 2025)	67.42	61.09	-	-
G2SERC (ours)	**70.42**	**67.64**	**40.45**	**59.64**

Table 2: Overall performance on the four datasets.

(a) IEMOCAP

Emotion	Sad.	Hap.	Exc.	Fru.	Ang.	Neu.	*Avg.*
SACL-LSTM	**84.78**	**56.91**	69.70	65.02	64.09	**70.00**	68.42
G2SERC	83.61	52.94	**73.45**	**67.98**	**68.02**	69.62	**69.27**
Improve	-1.17	-3.97	+3.75	+2.96	+3.93	-0.38	+0.85

(b) MELD

Emotion	Sad.	Fea.	Neu.	Joy.	Ang.	Sur.	Dis.	*Avg.*
SACL-LSTM	41.34	**26.23**	**80.17**	64.98	52.35	58.77	**31.47**	50.76
G2SERC	**43.30**	22.78	80.12	**65.70**	**55.09**	**61.49**	28.92	**51.06**
Improve	+1.96	-3.45	-0.05	+0.72	+2.74	+2.72	-2.55	+0.30

(c) EmoryNLP

Emotion	Joy.	Mad.	Pea.	Neu.	Sad.	Pow.	Sca.	*Avg.*
SACL-LSTM	**54.78**	37.68	11.66	**55.42**	**25.83**	5.43	37.11	32.56
G2SERC	54.04	**39.25**	**21.17**	52.75	22.38	**19.08**	**40.19**	**35.55**
Improve	-0.74	+1.57	+9.51	-2.67	-3.45	+13.67	+3.08	+2.99

(d) DailyDialogue

Emotion	Non.	Sur.	Hap.	Fea.	Ang.	Dis.	Sad.	*Avg.*
SACL-LSTM†	83.93	37.77	55.94	41.67	40.47	27.16	27.33	44.90
G2SERC	**91.97**	**52.36**	**65.01**	**51.85**	**42.93**	**32.43**	**38.89**	**53.63**
Improve	+8.04	+14.59	+9.07	+10.18	+2.46	+5.27	+11.56	+8.74

Table 3: Fine-grained performance comparison (on F1 score) between G2SERC and SACL-LSTM. † Results reproduced from the publicly available code.

Method	IEMOCAP	MELD	EmoryNLP	DailyDialogue
Full model	**70.42**	**67.64**	**40.45**	**59.64**
w/o Encoder	67.86	63.63	40.12	59.21
w/o GRU_s	67.27	63.62	39.12	59.32
w/o GRU_u	68.00	63.74	38.77	59.41

Table 4: Model ablation results.

integrating relation-aware graph encoding with sequential emotion decoding.

Table 3 reports the per-class performance comparison on the benchmark datasets. Fine-grained emotion recognition remains challenging for at least two reasons. First, some emotion categories are semantically similar and therefore hard to distinguish (e.g., *happy* versus *excited*, and *peaceful* versus *neutral*). Second, these benchmarks often exhibit long-tailed label distributions, where minority classes receive limited supervision. Among the compared baselines, SACL-LSTM achieves the strongest fine-grained results, likely benefiting from its contrastive learning objective that enhances intra-class discrimination. Under this setting, G2SERC remains competitive across most categories while consistently improving the average F1 by +0.85, +0.30, +2.99, and +8.74 on IEMOCAP, MELD, EmoryNLP, and DailyDialogue, respectively.

5.2 Ablation Study

Model Ablation. Table 4 reports model-level ablation results, examining the contributions of the graph encoder and the GRU-based decoder components. To remove the encoder, we replace it with a lightweight alternative: we obtain the global context vector $\boldsymbol{c}$ by pooling RoBERTa-based utterance representations, and derive speaker representations $\boldsymbol{H}^{(s)}$ by embedding one-hot speaker indicators. As shown in Table 4, removing the encoder consistently degrades performance across all datasets

Method	IEMOCAP	MELD	EmoryNLP	DailyDialogue
Full model	**70.42**	**67.64**	**40.45**	**59.64**
w/o $r1$	67.94	63.44	39.07	59.33
w/o $r2$	67.99	63.54	39.32	59.37
w/o $r3$	69.07	63.49	38.96	59.19

Table 5: Edge-type ablation results.

Correctness Condition	Accuracy (%)
Correct on u_{i+1} only	84.31
Correct on (u_i, u_{i+1})	39.22
Correct on (u_{i-1}, u_{i+1})	74.51
Correct on (u_{i-1}, u_i, u_{i+1})	31.37

Table 6: Accuracy breakdown over utterance triplets $\langle u_{i-1}, u_i, u_{i+1} \rangle$.

(e.g., by -2.56 on IEMOCAP and -4.01 on MELD), indicating that explicit structural encoding is crucial for ERC. To isolate the roles of the two GRU modules, we remove each component individually. Without the speaker-level GRU (GRU_s), the decoder reduces to an utterance-level GRU; speaker states are neither updated nor used for prediction, leading to consistent drops across datasets. Similarly, removing the utterance-level GRU (GRU_u) eliminates utterance-level temporal modeling; we directly feed utterance representations to GRU_s, which also results in noticeable degradation. Overall, both components contribute meaningfully to the final performance.

Edge-type Ablation. Table 5 presents ablations over relation types, including inter-speaker dependency ($r1$), intra-speaker dependency ($r2$), and global semantic similarity ($r3$). We ablate $r1$-$r3$ while keeping the utterance-speaker affiliation edges ($r4$) unchanged, since removing $t4$ would disconnect speaker information from the encoder. Overall, removing any of the three utterance-utterance relations consistently hurts performance, indicating that each relation contributes to the final performance. Among them, ablating $r1$ and $r2$ causes more noticeable degradation on IEMOCAP and MELD, whereas removing $r3$ leads to large decreases on EmoryNLP. These findings suggest that different relation types provide complementary signals for ERC.

5.3 Emotion Transition Analysis

To assess whether G2SERC maintains stable predictions after a local emotional deviation, we conduct an emotion transition analysis on dialogue segments from the IEMOCAP test set. For each conversation, we define the *dominant emotion* as the most frequent label across all utterances. We then extract all triplets $\langle u_{i-1}, u_i, u_{i+1} \rangle$ whose labels satisfy $y_{i-1} = y_{i+1} = y^*$ and $y_i \neq y^*$, where y^* denotes the dominant emotion of the dialogue. Here, the middle utterance u_i serves as a *deviation utterance* with respect to the global affective context. Our analysis focuses on whether such a deviation interferes with the prediction of the subsequent utterance u_{i+1}.

As shown in Table 6, when evaluated solely on u_{i+1}, G2SERC achieves an accuracy of 84.31% despite the deviation at u_i, indicating stable forward prediction under local emotional perturbations. When both u_i and u_{i+1} must be predicted correctly, the accuracy drops to 39.22%, reflecting the difficulty of modeling transient deviations. In contrast, requiring correctness on utterances aligned with the dominant emotion (u_{i-1} and u_{i+1}) yields a higher accuracy of 74.51%. Finally, correctly predicting the full triplet remains challenging (31.37%). Overall, these results suggest that the model tends to preserve global affective consistency even when short-lived local deviations occur.

6 Conclusion

We propose G2SERC, a graph-to-sequence framework that explicitly integrates dialogue-level affective context with utterance-level emotional dynamics for ERC. G2SERC constructs a speaker-aware heterogeneous graph with relation-specific edges to encode diverse conversational dependencies and to support structured context aggregation. On top of this graph, a multi-layer relation-aware encoder captures both global dialogue-level affective context and speaker-level affective states. We further propose a coupled dual-GRU decoder that separately tracks utterance- and speaker-level emotional dynamics, enabling context-conditioned emotion prediction along the dialogue flow. Extensive experiments on multiple benchmarks demonstrate that G2SERC is effective and competitive across datasets.

Limitations

Although the proposed graph-to-sequence framework G2SERC is effective for emotion recognition in conversation, it has several limitations. First, explicitly modeling heterogeneous conversational dependencies and multi-level emotional dynamics

requires a relatively complex architecture, including relation-specific graph encoding and a coupled recurrent decoder. This design increases both implementation complexity and computational overhead, which may hinder deployment in resource-constrained environments or latency-sensitive applications. Second, while G2SERC is conceptually modality-agnostic, extending it to multimodal ERC would require introducing additional node types and cross-modal relations. Such extensions would complicate graph construction and message passing, and may raise scalability challenges as the number of modalities increases. Addressing these issues—e.g., simplifying the architecture and improving scalability for multimodal settings—is an important direction for future work.

More broadly, emotion recognition in conversation is inherently subjective and context-dependent, and models trained on annotated datasets may inherit annotator and dataset biases. Therefore, G2SERC should be used as an assistive tool rather than as a definitive assessment of human emotional states.

References

Wei Ai, Fuchen Zhang, Yuntao Shou, Tao Meng, Haowen Chen, and Keqin Li. 2025. Revisiting multimodal emotion recognition in conversation from the perspective of graph spectrum. In *AAAI2025*.

Yinan Bao, Qianwen Ma, Lingwei Wei, Wei Zhou, and Songlin Hu. 2022. Speaker-guided encoder-decoder framework for emotion recognition in conversation. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence*, pages 4051–4057, Vienna, Austria. International Joint Conferences on Artificial Intelligence Organization.

Shaked Brody, Uri Alon, and Eran Yahav. 2022. How attentive are graph attention networks? In *International Conference on Learning Representations*.

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N. Chang, Sungbok Lee, and Shrikanth S. Narayanan. 2008. Iemocap: Interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, 42(4):335–359.

Feiyu Chen, Jie Shao, Shuyuan Zhu, and Heng Tao Shen. 2023. Multivariate, multi-frequency and multimodal: Rethinking graph neural networks for emotion recognition in conversation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10761–10770, Vancouver, BC, Canada. IEEE.

Yahui Fu, Shogo Okada, Longbiao Wang, Lili Guo, Yaodong Song, Jiaxing Liu, and Jianwu Dang. 2021. Consk-gcn: Conversational semantic- and knowledge-oriented graph convolutional network for multimodal emotion recognition. In *2021 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6.

Yahui Fu, Shogo Okada, Longbiao Wang, Lili Guo, Yaodong Song, Jiaxing Liu, and Jianwu Dang. 2022. Context- and knowledge-aware graph convolutional network for multimodal emotion recognition. *IEEE MultiMedia*, 29(3):91–100.

Yao Fu, Shaoyang Yuan, Chi Zhang, and Juan Cao. 2023. Emotion recognition in conversations: A survey focusing on context, speaker dependencies, and fusion methods. *Electronics*, 12(22):4714.

Deepanway Ghosal, Navonil Majumder, Alexander Gelbukh, Rada Mihalcea, and Soujanya Poria. 2020. Cosmic: Commonsense knowledge for emotion identification in conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2470–2481, Online. Association for Computational Linguistics.

Deepanway Ghosal, Navonil Majumder, Soujanya Poria, Niyati Chhaya, and Alexander Gelbukh. 2019. Dialoguegcn: A graph convolutional neural network for emotion recognition in conversation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 154–164, Hong Kong, China. Association for Computational Linguistics.

Ziwei Gong, Qingkai Min, and Yue Zhang. 2023. Eliciting rich positive emotions in dialogue generation. In *Proceedings of the First Workshop on Social Influence in Conversations (SICon 2023)*, pages 1–8, Toronto, Canada. Association for Computational Linguistics.

Zhinan Gou, Yuchen Long, Jieli Sun, and Kai Gao. 2025. Tg-erc: Utilizing three generation models to handle emotion recognition in conversation tasks. *Expert Systems with Applications*, 268:126269.

Devamanyu Hazarika, Soujanya Poria, Amir Zadeh, Erik Cambria, Louis-Philippe Morency, and Roger Zimmermann. 2018. Conversational memory network for emotion recognition in dyadic dialogue videos. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2122–2132, New Orleans, Louisiana. Association for Computational Linguistics.

Dou Hu, Yinan Bao, Lingwei Wei, Wei Zhou, and Songlin Hu. 2023. Supervised adversarial contrastive learning for emotion recognition in conversations. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1:*

Long Papers), pages 10835–10852, Toronto, Canada. Association for Computational Linguistics.

Dou Hu, Xiaolong Hou, Lingwei Wei, Lianxin Jiang, and Yang Mo. 2022a. 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)*, pages 7037–7041.

Dou Hu, Lingwei Wei, and Xiaoyong Huai. 2021a. Dialoguecrn: Contextual reasoning networks for emotion recognition in conversations. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7042–7052, Online. Association for Computational Linguistics.

Guimin Hu, Ting-En Lin, Yi Zhao, Guangming Lu, Yuchuan Wu, and Yongbin Li. 2022b. Unimse: Towards unified multimodal sentiment analysis and emotion recognition. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7837–7851, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jingwen Hu, Yuchen Liu, Jinming Zhao, and Qin Jin. 2021b. Mmgcn: Multimodal fusion via deep graph convolution network for emotion recognition in conversation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5666–5675, Online. Association for Computational Linguistics.

Wenxiang Jiao, Michael Lyu, and Irwin King. 2020a. Exploiting unsupervised data for emotion recognition in conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4839–4846, Online. Association for Computational Linguistics.

Wenxiang Jiao, Michael Lyu, and Irwin King. 2020b. Real-time emotion recognition via attention gated hierarchical memory network. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8002–8009.

Ran Jing, Geng Tu, Yice Zhang, and Ruifeng Xu. 2026. Causal-erc: A multimodal framework with causal prompting for emotion recognition in conversations with large language models. In *AAAI 2026*.

Tanmay Khule, Rishabh Agrawal, and Apurva Narayan. 2024. Pfa-erc: Psuedo-future augmented dynamic emotion recognition in conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 16196–16207, Miami, Florida, USA. Association for Computational Linguistics.

Taewoon Kim and Piek Vossen. 2021. Emoberta: Speaker-aware emotion recognition in conversation with roberta. *Preprint*, arXiv:2108.12009.

Akshi Kumar, Prakhar Dogra, and Vikrant Dabas. 2015. Emotion analysis of twitter using opinion mining. In *2015 Eighth International Conference on Contemporary Computing (IC3)*, pages 285–290.

Shanglin Lei, Guanting Dong, Xiaoping Wang, Keheng Wang, Runqi Qiao, and Sirui Wang. 2024. Instructerc: Reforming emotion recognition in conversation with multi-task retrieval-augmented large language models. *Preprint*, arXiv:2309.11911.

Jiang Li, Xiaoping Wang, Yingjian Liu, and Zhigang Zeng. 2024a. Cfn-esa: A cross-modal fusion network with emotion-shift awareness for dialogue emotion recognition. *IEEE Transactions on Affective Computing*, 15(4):1919–1933.

Jiang Li, Xiaoping Wang, Guoqing Lv, and Zhigang Zeng. 2023. Graphmft: A graph network based multimodal fusion technique for emotion recognition in conversation. *Neurocomputing*, 550(C).

Jiang Li, Xiaoping Wang, and Zhigang Zeng. 2025. Tracing intricate cues in dialogue: Joint graph structure and sentiment dynamics for multimodal emotion recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(10):8786–8803.

Jingye Li, Donghong Ji, Fei Li, Meishan Zhang, and Yijiang Liu. 2020. Hitrans: A transformer-based context- and speaker-sensitive model for emotion detection in conversations. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4190–4200, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Mingcheng Li, Dingkang Yang, Xiao Zhao, Shuaibing Wang, Yan Wang, Kun Yang, Mingyang Sun, Dongliang Kou, Ziyun Qian, and Lihua Zhang. 2024b. Correlation-decoupled knowledge distillation for multimodal sentiment analysis with incomplete modalities. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12458–12468, Seattle, WA, USA. IEEE.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. Dailydialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 986–995, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Zheng Lian, Fan Zhang, Yazhou Zhang, Jianhua Tao, Rui Liu, Haoyu Chen, and Xiaobai Li. 2025. Affectgpt-r1: Leveraging reinforcement learning for open-vocabulary multimodal emotion recognition. *Preprint*, arXiv:2508.01318.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *Preprint*, arXiv:1907.11692.

Navonil Majumder, Soujanya Poria, Devamanyu Hazarika, Rada Mihalcea, Alexander Gelbukh, and Erik Cambria. 2019. Dialoguernn: An attentive rnn for emotion detection in conversations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6818–6825.

Yuzhao Mao, Guang Liu, Xiaojie Wang, Weiguo Gao, and Xuan Li. 2021. Dialoguetrm: Exploring multi-modal emotional dynamics in a conversation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2694–2704, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Md Saef Ullah Miah, Md Mohsin Kabir, Talha Bin Sarwar, Mejdl Safran, Sultan Alfarhood, and M. F. Mridha. 2024. A multimodal approach to cross-lingual sentiment analysis with ensemble of transformer and llm. *Scientific Reports*, 14(1):9603.

Soujanya Poria, Erik Cambria, Devamanyu Hazarika, Navonil Majumder, Amir Zadeh, and Louis-Philippe Morency. 2017. Context-dependent sentiment analysis in user-generated videos. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 873–883, Vancouver, Canada. Association for Computational Linguistics.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2019a. Meld: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 527–536, Florence, Italy. Association for Computational Linguistics.

Soujanya Poria, Navonil Majumder, Rada Mihalcea, and Eduard Hovy. 2019b. Emotion recognition in conversation: Research challenges, datasets, and recent advances. *IEEE Access*, 7:100943–100953.

Neeraj Anand Sharma, A. B. M. Shawkat Ali, and Muhammad Ashad Kabir. 2025. A review of sentiment analysis: Tasks, applications, and deep learning techniques. *International Journal of Data Science and Analytics*, 19(3):351–388.

Weizhou Shen, Junqing Chen, Xiaojun Quan, and Zhixian Xie. 2021a. Dialogxl: All-in-one xlnet for multi-party conversation emotion recognition. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(15):13789–13797.

Weizhou Shen, Siyue Wu, Yunyi Yang, and Xiaojun Quan. 2021b. Directed acyclic graph network for conversational emotion recognition. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1551–1560, Online. Association for Computational Linguistics.

Tao Shi and Shao-Lun Huang. 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)*, pages 14752–14766, Toronto, Canada. Association for Computational Linguistics.

Yuntao Shou, Tao Meng, Wei Ai, and Keqin Li. 2025. Dynamic graph neural ode network for multi-modal emotion recognition in conversation. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 256–268, Abu Dhabi, UAE. Association for Computational Linguistics.

Rui Song, Fausto Giunchiglia, Lida Shi, Qiang Shen, and Hao Xu. 2023. Sunet: Speaker-utterance interaction graph neural network for emotion recognition in conversations. *Engineering Applications of Artificial Intelligence*.

Geng Tu, Tian Xie, Bin Liang, Hongpeng Wang, and Ruifeng Xu. 2024. Adaptive graph learning for multimodal conversational emotion detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19089–19097.

Yan Wang, Bo Wang, Yachao Zhao, Dongming Zhao, Xiaojia Jin, Jijun Zhang, Ruifang He, and Yuexian Hou. 2024. Emotion recognition in conversation via dynamic personality. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 5711–5722, Torino, Italia. ELRA and ICCL.

Yan Wang, Jiayu Zhang, Jun Ma, Shaojun Wang, and Jing Xiao. 2020. Contextualized emotion recognition in conversation as sequence tagging. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 186–195, 1st virtual meeting. Association for Computational Linguistics.

ChengYan Wu, Yiqiang Cai, Yang Liu, Pengxu Zhu, Yun Xue, Ziwei Gong, Julia Hirschberg, and Bolei Ma. 2025. Multimodal emotion recognition in conversations: A survey of methods, trends, challenges and prospects. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 6257–6274, Suzhou, China. Association for Computational Linguistics.

Sheng Wu, Jiaxing Liu, Longbiao Wang, Dongxiao He, Xiaobao Wang, and Jianwu Dang. 2024. Aimdit: Modality augmentation and interaction via multimodal dimension transformation for emotion recognition in conversations. *Preprint*, arXiv:2407.00743.

Shuo Yin and Guoqiang Zhong. 2024. Textgt: A double-view graph transformer on text for aspect-based sentiment analysis. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(17):19404–19412.

Fangxu Yu, Junjie Guo, Zhen Wu, and Xinyu Dai. 2024. Emotion-anchored contrastive learning framework

for emotion recognition in conversation. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 4521–4534, Mexico City, Mexico. Association for Computational Linguistics.

Sayyed M Zahiri and Jinho D Choi. 2017. Emotion detection on tv show transcripts with sequence-based convolutional neural networks. In *AAAI*.

Dong Zhang, Liangqing Wu, Changlong Sun, Shoushan Li, Qiaoming Zhu, and Guodong Zhou. 2019. Modeling both context- and speaker-sensitive dependence for emotion detection in multi-speaker conversations. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, pages 5415–5421, Macao, China. International Joint Conferences on Artificial Intelligence Organization.

Jinming Zhao, Tenggan Zhang, Jingwen Hu, Yuchen Liu, Qin Jin, Xinchao Wang, and Haizhou Li. 2022. M3ed: Multi-modal multi-scene multi-label emotional dialogue database. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5699–5710, Dublin, Ireland. Association for Computational Linguistics.

Peixiang Zhong, Di Wang, and Chunyan Miao. 2019. Knowledge-enriched transformer for emotion detection in textual conversations. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 165–176, Hong Kong, China. Association for Computational Linguistics.

Lixing Zhu, Gabriele Pergola, Lin Gui, Deyu Zhou, and Yulan He. 2021. Topic-driven and knowledge-aware transformer for dialogue emotion detection. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1571–1582, Online. Association for Computational Linguistics.

A Datasets

We evaluate our model on four widely used benchmark datasets for emotion recognition in conversation, namely IEMOCAP, MELD, DailyDialog, and EmoryNLP. The statistics of these datasets are summarized in Table 1.

IEMOCAP (Busso et al., 2008) is a multimodal ERC dataset consisting of dyadic conversations performed by professional actors following scripted scenarios. Each utterance is annotated with one of six emotion categories: neutral, happiness, sadness, anger, frustrated, and excited.

MELD (Poria et al., 2019a) is a multimodal conversational emotion dataset collected from the TV series *Friends*. It contains multi-party conversations annotated with seven emotion labels: neutral, happiness, surprise, sadness, anger, disgust, and fear.

EmoryNLP (Zahiri and Choi, 2017) is another ERC dataset derived from TV show scripts of *Friends*, differing from MELD in both scene selection and emotion annotation scheme. It includes seven emotion categories: neutral, sad, mad, scared, powerful, peaceful, and joyful.

DailyDialog (Li et al., 2017) is a large-scale text-based dataset composed of human-written daily conversations. Each utterance is labeled with one of seven emotion categories: neutral, happiness, surprise, sadness, anger, disgust, and fear. Since explicit speaker annotations are not available, we treat utterance turns as speaker turns by default.

B Baseline Methods

DialogueRNN (Majumder et al., 2019) models conversational emotion dynamics using recurrent neural networks with explicit speaker state tracking.

SGED (Bao et al., 2022) enhances sequence-based emotion recognition by modeling speaker-aware emotional dynamics with gated recurrent architectures.

SACL-LSTM (Hu et al., 2023) incorporates supervised contrastive learning into an LSTM-based framework to improve emotion representation learning.

DialogXL (Shen et al., 2021a) extends Transformer architectures to conversational settings by modeling long-range contextual dependencies across dialogue turns.

MultiEMO (Shi and Huang, 2023) adopts a Transformer-based framework to capture multimodal emotional cues through cross-modal attention mechanisms.

CFN-ESA (Li et al., 2024a) introduces emotion state augmentation within a Transformer framework to enhance emotion representation in conversations.

DialogueGCN (Ghosal et al., 2019) formulates emotion recognition in conversation as a graph learning problem, explicitly modeling speaker interactions and contextual dependencies.

DAG-ERC (Shen et al., 2021b) employs directed acyclic graphs to capture causal and temporal dependencies among utterances in a conversation.

GS-MCC (Ai et al., 2025) leverages graph-based multimodal contextual modeling to improve con-

Para.	IEMOCAP	MELD	EmoryNLP	DailyDialog
Optimizer	AdamW			
Dropout rate	0.24	0.10	0.20	0.30
GNN layers	2			
Hidden dim	1024			
Window size	1			
Learning rate	5×10^{-5}			
Weight decay	1×10^{-2}			
Batch size	16	64	32	64
Epochs	150	80	60	60
Seed	3407			

Table 7: Hyperparameter settings.

versational emotion recognition.

EmoBERTa (Kim and Vossen, 2021) fine-tunes pretrained language models for utterance-level emotion classification without explicit conversational structure modeling.

ERC-DP (Wang et al., 2024) incorporates dynamic personality representations into a pretrained language model framework for emotion recognition in conversation.

AffectGPT-R1 (Lian et al., 2025) adopts a prompt-based paradigm with pretrained language models to perform emotion recognition in conversational contexts.

C Hyperparameter settings.

This section summarizes the hyperparameter settings used in all experiments, as listed in Table 7.

D Discussion on Potential Risks

While our study does not introduce new datasets or real-world deployments, emotion predictions may be misinterpreted if used without proper contextual understanding. We emphasize that the proposed framework is intended for research purposes and should not be directly applied to high-stakes decision-making scenarios without further validation.