

Article

Multimodal Emotion Recognition in Conversations Using Transformer and Graph Neural Networks

Hua Jin , Tian Yang, Letian Yan, Changda Wang  and Xuehua Song 

School of Computer Science and Computer Science and Communication Engineering, Jiangsu University, Zhenjiang 212013, China; jinhua@ujs.edu.cn (H.J.); yangtian8879@163.com (T.Y.); 2222408011@stmail.ujs.edu.cn (L.Y.); changda@ujs.edu.cn (C.W.)

* Correspondence: songxh@ujs.edu.cn

Abstract

To comprehensively capture conversational emotion information within and between modalities, address the challenge of global and local feature modelling in conversation, and enhance the accuracy of multimodal conversation emotion recognition, we present a model called Multimodal Transformer and GNN for Emotion Recognition in Conversations (MTG-ERC). The model incorporates a multi-level Transformer fusion module that employs multi-head self-attention and cross-modal attention mechanisms to effectively capture interaction patterns within and between modalities. To address the shortcomings of attention-mechanism-based models in capturing short-term dependencies, we introduce a directed multi-relational graph fusion module, which employs directed graphs and multiple relation types to achieve efficient multimodal information fusion and to model short-term, speaker-dependent emotional shifts. By integrating the outputs of these two modules, the MTG-ERC model effectively combines global and local conversational emotion features and enhances intra-modal and inter-modal emotional interactions. The proposed model shows consistent improvements (around 1% absolute) in both accuracy and weighted F1 on the IEMOCAP and MELD datasets when compared with other baseline models. This highlights the model's strong performance indicators and validates its effectiveness in comparison to existing models.



Academic Editor: Giorgio De Nunzio

Received: 29 August 2025

Revised: 31 October 2025

Accepted: 8 November 2025

Published: 11 November 2025

Citation: Jin, H.; Yang, T.; Yan, L.; Wang, C.; Song, X. Multimodal Emotion Recognition in Conversations Using Transformer and Graph Neural Networks. *Appl. Sci.* **2025**, *15*, 11971. <https://doi.org/10.3390/app152211971>

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

Keywords: emotion recognition in conversation; multimodality; attention mechanism; graph neural network

1. Introduction

In the context of the information age, a variety of conversational scenarios have become an integral aspect of our lives, giving rise to the emerging field of Emotion Recognition in Conversations (ERC). ERC is concerned with the identification of subtle emotional states within the context of conversational interactions, representing a crucial aspect of emotional computing. The model performs emotion recognition for each utterance based on the context of the conversation and the information about the speaker, employing a multi-category classification approach. In particular, this entails accurately classifying each utterance in multiple predefined emotion categories (such as joy, anger, sadness, surprise, etc.), rather than just a simple emotional positive and negative polarity analysis. This enables the model to capture complex emotional changes.

ERC plays a crucial role in various application areas, including conversation emotion monitoring, social media sentiment analysis and emotion perception in dialogue systems.

Additionally, ERC also shows potential application in recommendation systems, customer service feedback and healthcare. At the same time, ERC systems must operate under practical constraints such as latency, interpretability, and privacy when deployed in real-world interactive settings.

The most commonly used methods for modelling conversation context can be categorized into three groups: those based on recurrent neural networks, those based on attention mechanisms, and those based on graph neural networks [1]. Recurrent neural networks process sequence data in a stepwise manner, retaining the order information of the time series, but may suffer from inefficiency and vanishing/exploding gradients for long sequences. Methods based on attention mechanisms, exemplified by Transformer [2], capture global context through self-attention mechanisms. However, they may under-emphasize fine-grained, short-term, speaker-dependent emotional shifts in local conversational windows. Graph neural networks (GNNs) are capable of effectively capturing detailed local emotional information by collecting contextual information around the discourse. However, GNNs can find it more difficult to efficiently capture long-range, cross-modal dependencies at the conversation level.

The field of multimodal conversation emotion recognition currently faces two core challenges. The first is how to capture the information interactions within and between modes of discourse. The second is how to simultaneously acquire global and local conversation emotion features. Recent research has demonstrated the efficacy of attention mechanisms in multimodal fusion, facilitating the capture of pertinent information across modalities. However, the significance of emotional data within the modality itself is overlooked. To this end, we present a multi-level Transformer fusion module, which employs multi-head self-attention and cross-modal attention mechanisms to effectively capture information interactions within and between modalities, thereby enriching emotional information. To better capture speaker-conditioned short-term dependencies and complement global attention with local reasoning, we further introduce a directed multi-relational graph fusion module, which constructs a directed conversational graph with multiple relation types and fuses multimodal information within a bounded local context. By integrating these two branches, the model jointly models global long-range dependencies and local short-range emotional cues.

In summary, the main contributions of this work are as follows:

1. We design and present the MTG-ERC multimodal conversational emotion recognition model. MTG-ERC consists of a multi-level Transformer fusion module and a directed multi-relational graph fusion module. These two modules are combined through a gated fusion mechanism to integrate global (long-range, cross-modal) and local (short-term, speaker-aware) emotional features. MTG-ERC yields consistent improvements (approximately 1% absolute in both accuracy and weighted F1 on IEMOCAP and MELD) over a strong graph-based baseline, and we discuss the statistical significance of these gains rather than claiming large margins.
2. We introduce a multi-level Transformer fusion module. This module uses multi-head self-attention (intra-modal) together with cross-modal attention (inter-modal) to capture interactions within and between modalities and to model long-range conversational context. This branch focuses on global dependency modelling across the dialogue.
3. We introduce a directed multi-relational graph fusion module. This module represents the conversation as a directed graph with multiple explicitly defined relation types (e.g., intra-utterance cross-modal links, self-loops, temporally directed edges). It aggregates short-range contextual signals and speaker-dependent emotional shifts, enabling fine-grained local reasoning that complements the Transformer branch.

4. We conduct performance evaluation and ablation studies on the public IEMOCAP [3] and MELD [4]. We compare MTG-ERC against strong recent baselines, report averaged results across multiple runs, and analyze variance and paired significance to assess robustness. We also analyze the computational characteristics of the two modules to clarify the accuracy–efficiency trade-off and guide deployment considerations.

2. Related Works

2.1. Multimodal Information Fusion

Deep learning has given rise to a significant body of research on multimodal methods. In 2017, Zadeh et al. [5] put forth the TFN model, which simulates the interaction of unimodal, bimodal, and trimodal modes through the calculation of Cartesian products. Nevertheless, the tensor fusion of TFN necessitates a considerable number of parameters, which consequently increases the computational complexity and memory consumption. To address this issue, Liu et al. [6] proposed the LMF model in 2018, which employs low-rank tensors for multimodal fusion, effectively reducing the model complexity and the number of parameters. Despite the success of these methods in multimodal fusion, they have not yet addressed the issue of temporal alignment across different modalities. To address this challenge, Tsai et al. [7] proposed the MulT model in 2019, which employs a pairwise cross-modal attention mechanism to capture the interaction between different time points and different modalities. However, this approach may under-represent the fine-grained information within each individual modality. In 2020, Sahay et al. [8] proposed the LMF-MulT model, which combined the advantages of MulT and LMF. It uses LMF to obtain multimodal representation and realizes the interaction fusion between modalities through a cross-modal attention mechanism. However, since LMF uses a low-rank matrix representation, some high-dimensional information may be lost, thereby making it harder to retain very subtle affective cues and complex interactions.

Building on these developments, we introduce a multi-level Transformer fusion module (referred to in this work as the Multi-Level Transformer Fusion Module, MTFM). This module combines intra-modal self-attention and inter-modal cross-attention to model both within-modality and cross-modality interactions. In this way, it captures long-range conversational context and enriches the emotional information present in each modality without relying solely on early fusion or purely pairwise cross-modal attention.

2.2. Conversational Emotion Recognition

The ERC model improves sentiment classification by modelling contextual information. There are three types of ERC model: those based on recurrent neural networks, those based on attention mechanisms, and those based on graph neural networks.

In 2019, Majumder et al. [9] proposed the DialogueRNN model, which is based on recurrent neural networks. This model employs three types of GRU to update the speaker features, context, and emotions in the conversation, respectively, thereby effectively capturing the dependencies between speakers and the conversation. Subsequently, Ghosal et al. [10] proposed the COSMIC model in 2020, which incorporates external common-sense knowledge into the DialogueRNN to enhance model performance. The model's ability to comprehend emotions is enhanced through the introduction of an external knowledge base; however, this reliance on external sources may be a limitation. In 2021, Hu et al. [11] further introduced a multi-round reasoning module on the bidirectional LSTM to model emotions from a cognitive perspective. Nevertheless, models based on recurrent neural networks may suffer from vanishing or exploding gradients and reduced efficiency when modelling very long conversational dependencies.

Among the models based on the attention mechanism, Shen et al. [12] proposed the DialogXL model in 2021 with the objective of obtaining longer-term contextual information in the field of conversation emotion recognition. In 2021, Kim et al. [13] proposed the EmoBERTa model, which incorporates speaker information to facilitate the characteristics of speakers in the conversation and their relationships, and utilizes this information to enhance the accuracy of ERC. In 2022, Li et al. [14] proposed the CoG-BART model. The system conducts supervised contrastive learning to improve the model's capacity for processing contextual information. However, it should be noted that a substantial amount of labelled data is required for supervised learning. Hu et al. [15] also proposed the Unimse model in 2022 with the objective of unifying the tasks of Multimodal Sentiment Analysis (MSA) and ERC. The model conducts modal fusion at both the syntactic and semantic levels, while also implementing contrastive learning between modalities and samples, with the aim of better capturing the differences and consistency between emotions. However, purely attention-based models may under-emphasize short-term, speaker-dependent emotional shifts in local conversational windows.

Among the models based on graph neural networks, Ghosal et al. [16] proposed the DialogueGCN model in 2019. This model regards each dialogue as a graph, associates each utterance with the surrounding utterances, and uses the graph structure to capture the complex relationship between utterances. In 2020, Ishiwatari et al. [17] proposed the RGAT model, which incorporates position encoding into DialogueGCN and combines positional information to enhance the accuracy of ERC. In 2021, Hu et al. [18] proposed the MMGCN model, which extends the DialogueGCN model to encompass multimodal scenarios. It performs graph convolution operations on the utterance representations of different modalities, thereby effectively capturing the interaction between different modalities. In 2023, Wang et al. [19] proposed a GraphCFC module. This approach effectively models context and interaction information; however, it does not explicitly parameterize or analyze the full set of directed relational edge types in the conversation graph, which may limit fine-grained relational reasoning. More recent work (2023–2025) has begun to explore hybrid architectures that integrate Transformer-style global reasoning with graph-based local structure for conversational emotion recognition, highlighting the need to jointly model both long-range dependencies and local speaker-aware cues.

In this work, we introduce a directed multi-relational graph fusion module (DMRGF), which represents each conversation as a directed graph with multiple explicitly defined relation types (including intra-utterance cross-modal links, self-loops, and temporally directed edges). This module performs localized message passing to capture short-term, speaker-conditioned emotional dynamics. When combined with the multi-level Transformer fusion module described above, the overall MTG-ERC framework integrates global (long-range, cross-modal) and local (short-range, speaker-aware) conversational emotion features. We emphasize this integration as a reproducible formulation rather than claiming an entirely new paradigm.

3. Methodology

This paper presents a multimodal conversational emotion recognition model named MTG-ERC, which is composed of four main modules:

- (1) The unimodal feature extraction module is responsible for extraction of features pertaining to each modality. The text features are extracted using a Transformer network, while the audio and visual features are extracted through a fully connected network.
- (2) Multi-level Transformer fusion module: The use of intra-modal and inter-modal Transformer networks enables the capture of interactive information in speech sequences, facilitating the integration of intra-modal and inter-modal interactions.

- (3) Directed Multi-Lateral Relationship Graph Fusion and Module: The construction of a directed graph of multi-variable relationships enables the learning of local context representation, the capture of the relationship between discourse and modality, and the enhancement of the expression of local features.
- (4) Multimodal emotion classification module: The weights of each modality are adjusted dynamically using the softmax function. The global and local context features obtained from the multi-level Transformer fusion module and the directed multi-lateral relationship graph fusion and module are spliced, and comprehensive integration of global and local conversational emotion features is achieved. The emotion label of each sentence is then output.

The comprehensive framework of the model is illustrated in Figure 1.

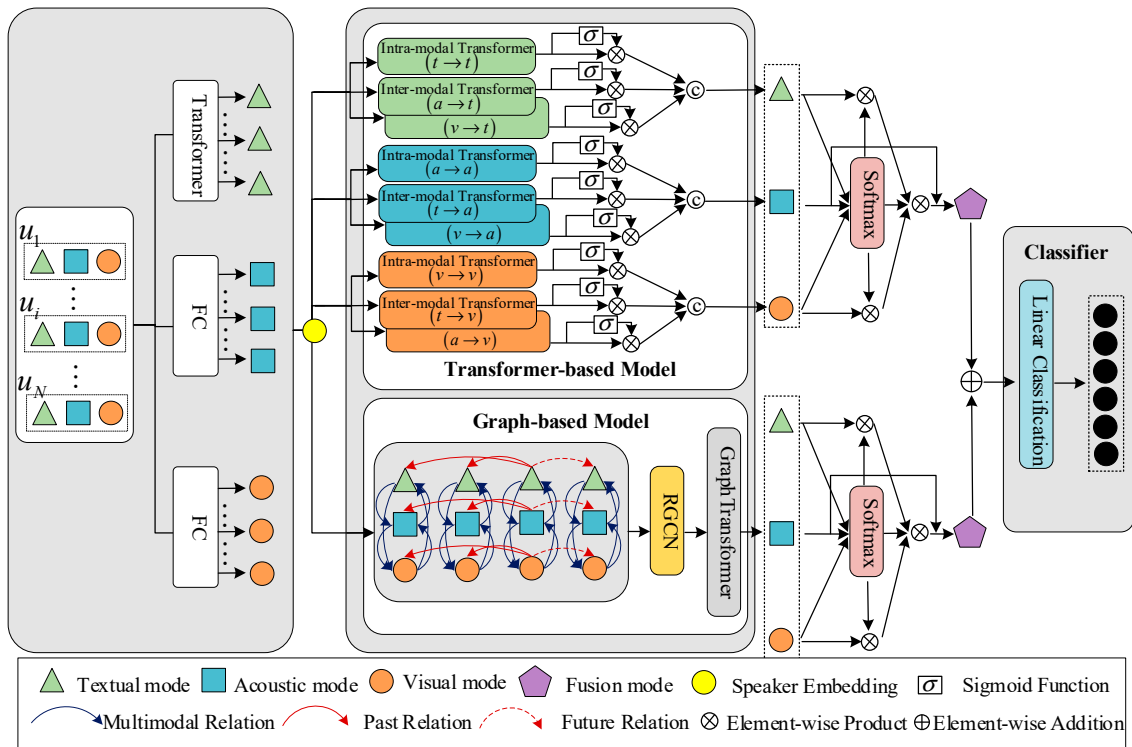


Figure 1. MTG-ERC model.

3.1. Task Definition

A conversation is constituted by a series of consecutive utterances $\{u_1, u_2, \dots, u_N\}$ and speakers $\{s_1, s_2, \dots, s_M\}$. Each utterance u_i comprises three modalities: text, audio, and vision, represented as u_i^t , u_i^a , and u_i^v . The objective of the ERC task is to predict the label denoting the emotional state of the speaker is assigned to each utterance, based on a predefined set of emotion labels. $Y = [y_1, y_2, \dots, y_K]$.

3.2. Feature Extraction

In the context of discourse u_i , each data modality evinces a distinct emotional proclivity. The discourse representation is derived from the original text, audio, and visual modal information $u_i^t \in R^{d_t}$, $u_i^a \in R^{d_a}$, $u_i^v \in R^{d_v}$, wherein d_t , d_a , d_v represent the feature dimensions of each modality.

In the case of textual modality, the process of textual feature extraction is carried out using the Transformer model in order to extract the semantic features x_i^t from u_i^t :

$$x_i^t = \text{Transformer}(u_i^t, W_{trans}^t) \quad (1)$$

where W_{trans}^l is the Transformer parameter to be learned.

In the case of audio and visual modalities, fully connected networks are employed for the purpose of feature extraction.

$$\begin{aligned} x_i^a &= FC(u_i^a; W_{fc}^a) \\ x_i^v &= FC(u_i^v; W_{fc}^v) \end{aligned} \quad (2)$$

where W_{fc}^a and W_{fc}^v are trainable parameters in the fully connected layer.

3.3. Speaker Embedding

To capture the speaker information in the speech sequence, a speaker embedding module is designed to enhance the speaker information. This involves mapping the speaker in the conversation into a vector, which more accurately reflects the speaker's characteristics.

$$S_{emb} = \text{Embedding}(S, M) \quad (3)$$

where S represents the set of speakers, M represents the total number of participants in the conversation, and S_{emb} represents the embedding vector matrix of the speakers.

The extracted utterance features can be further enhanced by the addition of the corresponding speaker embeddings:

$$X_\tau = \eta S_{emb} + x_\tau, \tau \in \{l, a, v\} \quad (4)$$

where x_τ respectively represent the global context representation of the entire conversation after feature extraction of text, audio, and visual modalities, X_τ represents the global context representation of each modality after speaker embedding enhancement, and $\eta \in [0, 1]$ represents the contribution rate.

3.4. Multi-Level Transformer Fusion Module

Intra-Modal and Inter-Modal Transformers

Intra-modal and Inter-modal Transformer networks have been developed with the objective of simulating the interaction within and between modalities in multimodal discourse sequences. These networks are designated as Intra-Transformer and Inter-Transformer, respectively, and their structural configurations are illustrated in Figure 2.

Previous research has concentrated on the efficient interaction with the emotional aspects of each modality, with little attention paid to the emotional characteristics of the modality itself. This paper presents the Intra-Transformer network, which employs the multi-head self-attention mechanism to capture long-distance dependencies in the feature data, thus enhancing the emotional feature representation of each modality.

The Intra-Transformer network takes the global context representation X_τ of each modality as input Q, K, V , captures the intra-modal interactions between discourse sequences, and enhances the emotional representation of discourse.

$$X_{\tau \rightarrow \tau} = \text{Intra-Transformer}(X_\tau, X_\tau, X_\tau) \quad (5)$$

In the formula, $\tau \in \{l, a, v\}$.

Although the data of different modalities are heterogeneous, each modality contains the sentiment information of the samples. Building on the MulT model, this paper presents an Inter-Transformer network, which combines multimodal feature data in pairs and inputs them into the cross-modal attention network. Additionally, the network employs

auxiliary modal data to cyclically reinforce the target modal data, thereby supplementing the sentiment information in the target modal features.

The Inter-Transformer network orders $Q = X_\tau, K = V = X_v$, so that the τ modality obtains the information of the v modality, thereby capturing the interaction between the various modalities.

$$X_{v \rightarrow \tau} = \text{Inter-Transformer}(X_\tau, X_v, X_v) \quad (6)$$

In the formula, $\tau \in \{l, a, v\}, v \in \{l, a, v\} - \{\tau\}$.

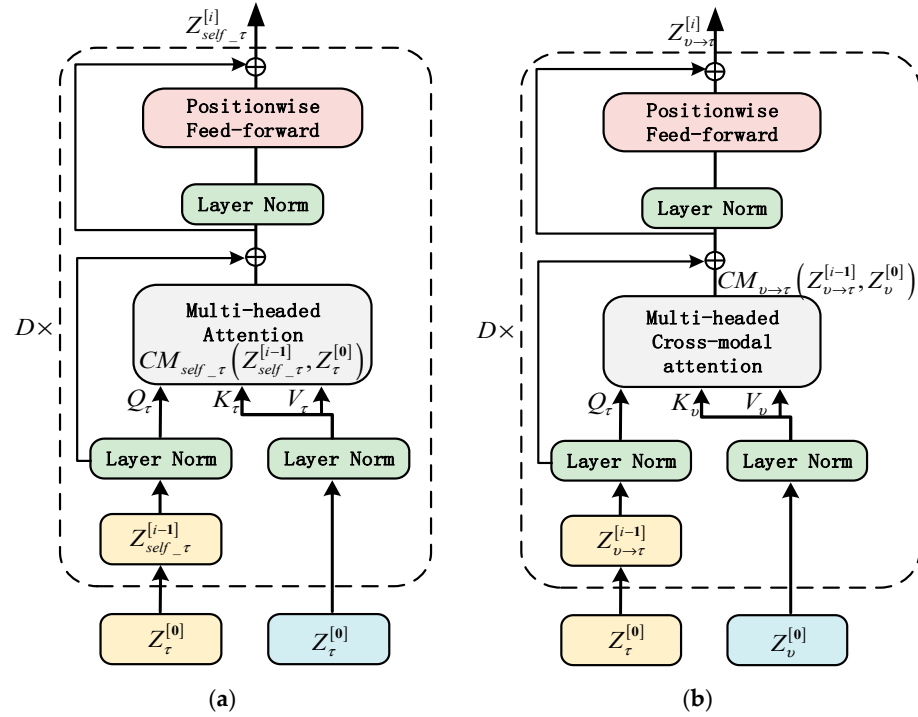


Figure 2. Intra-Transformer structure (a) and Inter-Transformer structure (b).

3.5. Single-Modality Gated Fusion

The utilization of a gated neural network, through the gating mechanism, enables the filtration of unimportant information in $X_{v \rightarrow \tau}$, while retaining the useful information.

$$g_{v \rightarrow \tau} = \sigma(W_{v \rightarrow \tau} \cdot X_{v \rightarrow \tau}) \quad (7)$$

$$X'_{v \rightarrow \tau} = X_{v \rightarrow \tau} \otimes g_{v \rightarrow \tau} \quad (8)$$

where $W_{v \rightarrow \tau}$ is the weight matrix, σ is the sigmoid function, and $g_{v \rightarrow \tau}$ is the gated neural network.

$X'_{\tau \rightarrow \tau}$, $X'_{v_1 \rightarrow \tau}$, and $X'_{v_2 \rightarrow \tau}$ are concatenated together and passed through a fully connected layer to generate enhanced sequence representations of text, audio, and visual modalities.

$$X'_\tau = W_\tau \cdot [X'_{\tau \rightarrow \tau}; X'_{v_1 \rightarrow \tau}; X'_{v_2 \rightarrow \tau}] + b_m \quad (9)$$

where W_τ and b_m are trainable parameters.

Let $X'_\tau = \{x'_{\tau 1}; x'_{\tau 2}; \dots; x'_{\tau N}\}$, where $x'_{\tau i}$ is the τ -modal representation of sentence u_i .

3.6. Directed Multilateral Graph Fusion Module

3.6.1. Graph Structure

The network structure of a conversation with N utterances can be represented by a directed graph $G = (v, \varepsilon, r)$, which represents the points, edges and relationships for text,

audio and vision. In order to better elucidate the discourse relationship between and within modalities, this module treats each utterance of each modality as a node of the graph, employing directed lines to connect past utterances and future utterances.

1. Node division: Each sentence u_i generates three modal nodes, denoted as u_i^l , u_i^a , and u_i^v . The feature vectors of the nodes are represented by x_i^l and x_i^a , respectively.
2. Edge division: The edges included in the edge set are contingent upon the scope of the context content considered during the modelling process. In a conversation, if the context content of each utterance considers all other utterances in the conversation, a fully connected graph will be generated, in which each vertex has an edge with all other vertices (including itself). However, this situation requires a significant amount of computing time, which is why this paper only considers local neighbour information and uses the context window $[P, F]$ to limit the scope when collecting local neighbour information. The parameters P and F represent the number of sentences connected to the past and future, respectively. In our implementation, we use an asymmetric context window $[P, F] = [9, 5]$ $[P, F] = [9, 5]$ $[P, F] = [9, 5]$. For each utterance, utu_tut , we create directed edges to up to 9 preceding utterances and 5 following utterances. We adopt an asymmetric window because emotional state is typically more influenced by accumulated past context (speaker history, escalation), while only short-range future reactions are informative. We also compared symmetric windows such as $[5, 5]$ and wider windows such as $[9, 9]; [9, 5]$ provided the best balance between classification performance and graph size, so we use this setting in all experiments.

The edge set is represented as $(u_i^\tau, u_j^\tau, r_{ij}) \in \varepsilon, \tau \in \{l, a, v\}$, which represents the interaction between points u_i^τ, u_j^τ and relationship type $r_{ij} \in R$. This module will consider two sets of relationships: R_{intra} and R_{inter} . Among them, R_{intra} represents the internal connection between the three modes in the same sentence, reflecting the interaction of multiple modes. In addition, each mode has a self-connection to strengthen the information of the mode itself. Formula (10) represents 9 types of edge relationships of R_{intra} .

$$R_{intra} = \left\{ \begin{array}{l} \left\{ (u_i^a, u_i^l), (u_i^v, u_i^l), (u_i^l, u_i^l) \right\} \\ \left\{ (u_i^l, u_i^a), (u_i^v, u_i^a), (u_i^a, u_i^a) \right\} \\ \left\{ (u_i^l, u_i^v), (u_i^a, u_i^v), (u_i^v, u_i^v) \right\} \end{array} \right\} \quad (10)$$

R_{inter} represents the impact of the context (including past and future discourse) of the same modal discourse on the current discourse, reflecting the interaction between different time nodes. The specific formula is the following:

$$R_{inter} = \left\{ \begin{array}{l} \left\{ (u_j \rightarrow u_i)^\tau | i - P < j < i \right\} \\ \left\{ (u_i \leftarrow u_j)^\tau | i < j < i + F \right\} \end{array} \right\} \quad (11)$$

where $\rightarrow$ and $\leftarrow$ represent the connection relationship with the past and the future, respectively, P and F represent the number of past and future related sentences.

3.6.2. Graph Learning

In order to differentiate between various types of edges and nodes and facilitate the capture of more nuanced structural information and feature representation, this module employs the Relational Graph Convolutional Network (RGCN) [20]. The node representation for each relationship type $r \in R$ is inferred using a mapping function $f(H, W_r)$ with a weighting matrix W_r . The final node representation $\sum_r^R f(H, W_r)$ can be calculated by combining 15 different edge types.

The formula for the i -th utterance is:

$$g_i^\tau = \sum_{r \in R} \sum_{j \in N_r(i)} \frac{1}{|N_r(i)|} W_r \cdot x_i^\tau + W_0 x_i^\tau \quad (12)$$

where $N_r(i)$ represents all neighbour nodes that have relationship r with node i . W_r and W_0 are learnable parameters in the RGCN.

In order to further extract information from the graph structure data, this module employs the Graph Transformer network [21]. The Graph Transformer network comprises layers containing a self-attention mechanism, which enables each node to attend to the information of other nodes in the graph, thereby enhancing the feature expression ability.

The g_i^τ obtained from the RGCN is transformed into the Graph Transformer network as follows:

$$o_i^\tau = \parallel_{c=1}^C \left[W_1 g_i^\tau + \sum_{i \in N(i)} a_{i,j}^\tau W_2 g_j^\tau \right] \quad (13)$$

where $N_r(i)$ is the set of nodes adjacent to node i , W_1 and W_2 are learnable parameters in the Graph Transformer network. C is the number of connected multi-head attentions.

The formula for the attention coefficient $a_{i,j}^\tau$ of node j is as follows:

$$a_{i,j}^\tau = \text{softmax} \left(\frac{(W_3 g_i^\tau)^T (W_4 g_j^\tau)}{\sqrt{d}} \right) \quad (14)$$

where W_3 and W_4 are learnable parameters.

After aggregating the entire graph, we get a new vector representation:

$$G^\tau = \{o_1^\tau, o_2^\tau, \dots, o_N^\tau\} \quad (15)$$

3.7. Multimodal Sentiment Classification

3.7.1. Multimodal Gated Fusion

A gating mechanism utilizing a Softmax function is devised to facilitate the dynamic learning of the modal weight associated with each utterance. In particular, the final multimodal representation of an utterance is determined by the following Equation (16):

$$\begin{aligned} [g_{ti}; g_{ai}; g_{vi}] &= \text{softmax}([W \cdot x'_{ti}; W \cdot x'_{ai}; W \cdot x'_{vi}]) \\ [g_{ti}; g_{ai}; g_{vi}] &= \text{softmax}([W \cdot o'_{ti}; W \cdot o'_{ai}; W \cdot o'_{vi}]) \end{aligned} \quad (16)$$

where W is the weight matrix, g_{ti} , g_{ai} , g_{vi} represent the weights learned by text, audio, and visual modalities, respectively.

The ultimate multimodal representation of each utterance is produced through the weighted addition of the enhanced representations of each modality.

$$\begin{aligned} x''_i &= \sum_{\tau \in \{t,a,v\}} x''_{\tau i} \otimes g_{\tau i} \\ o'_i &= \sum_{\tau \in \{t,a,v\}} o'_{\tau i} \otimes g_{\tau i} \end{aligned} \quad (17)$$

where x''_i and o'_i are the multimodal representations of utterance u_i in the Transformer module and the graph module, respectively.

The multimodal sequence generated by the multi-level Transformer fusion module and the directed multi-edge graph fusion module is r is shown like this:

$$X''_\tau = \{x''_{\tau 1}; x''_{\tau 2}; \dots; x''_{\tau N}\} \quad (18)$$

$$G'_\tau = \{o'_{\tau 1}, o'_{\tau 2}, \dots, o'_{\tau N}\} \quad (19)$$

3.7.2. Emotion Label Generation

The global and local context representations generated by the multi-level Transformer fusion module and the directed multilateral graph fusion module are concatenated together, as illustrated in Equation (20):

$$\begin{aligned} H &= \text{Fusion}([X, G]) \\ &= [\{x'_{\tau 1}, x'_{\tau 2}, \dots, x'_{\tau N}\}, \{o'_{\tau 1}, o'_{\tau 2}, \dots, o'_{\tau N}\}] \end{aligned} \quad (20)$$

where *Fusion* Indicates the connection method.

Input H into a fully connected layer, perform dimensionality reduction, and then make prediction and classification. This will allow us to ascertain the emotional label of speech u_i . The relevant calculation formula is shown in Equations (21)–(23):

$$p_i = \text{ReLU}(\phi_0 h_i + b_0) \quad (21)$$

$$q_i = \text{softmax}(\phi_1 p_i + b_1) \quad (22)$$

$$\hat{y}_i = \text{argmax}(q_i) \quad (23)$$

where ϕ_0 and ϕ_1 are learnable parameters.

4. Experiments

4.1. Dataset

The MTG-ERC model is evaluated using two datasets: IEMOCAP and MELD. The IEMOCAP dataset comprises conversational data from ten actors engaged in emotional interactions. The emotion labels are divided into six categories. This paper employs two experimental settings: one is a 4-category emotional scale (angry, sad, happy, neutral), and the other is a 6-category scale. The MELD dataset comprises 1432 dialogues from the TV series “Friends”. The data forms include video, text, and voice. The emotion labels are divided into seven categories: natural, surprised, fearful, sad, happy, disgusted, and angry. Table 1 illustrates the distribution of training samples and test samples in the two datasets.

Table 1. Distribution of dialogues and utterances.

Datasets	Dialogues		Utterances	
	Train	Test	Train	Test
IEMOCAP (6-way)	120	31	5810	1623
IEMOCAP (4-way)	120	31	3600	943
MELD	1153	280	11098	2610

4.2. Experimental Parameter Settings and Evaluation Indicators

In the IEMOCAP dataset, text features are extracted by sBERT with a feature dimension of 768, audio features are extracted by OpenSmile with a feature dimension of 100 and visual features are extracted by OpenFace with a feature dimension of 512.

In the MELD dataset, text features are extracted by the RoBERTa Large model with a feature dimension of 1024, audio features are extracted by OpenSmile with a feature dimension of 300 and visual features are extracted by DenseNet with a feature dimension of 342.

The MTG-ERC model selects Adam as the optimizer, the learning rate is set to 0.0003, the dropout rate is set to 0.5 and the number of multi-head attention used in the Graph

Transformer and Intra/Inter-Transformer networks is 7 and 2, respectively. The $[P, F]$ value is [9,5].

The model employs two evaluation indicators: accuracy (ACC) and weighted F1 value (w-F1). The weighted F1 value considers the weight of each category within the dataset, with the weight proportional to the number of utterances associated with a specific emotion tag. Given the uneven distribution of emotion samples within the dataset, this approach allows for a more accurate reflection of the model's overall performance. The calculation method is illustrated in Formula (25).

We set the learning rate to 0.0003 and the dropout rate to 0.5. These values were selected based on validation performance: we performed a small grid search over learning rates {0.0001, 0.0003, 0.0005} and dropout rates {0.3, 0.5, 0.7} on the development split of IEMOCAP, and (0.0003, 0.5) provided the best trade-off between convergence speed and generalization. We keep the same configuration for MELD for consistency and reproducibility.

$$wF_1 = \frac{\sum_{k=1}^K N_k \times F1_k}{\sum_{k=1}^K N_k} \quad (24)$$

4.3. Comparison with Baselines

This study compares the proposed MTG-ERC model with the baseline method in the multimodal emotion classification setting of the IEMOCAP dataset. In the table, bold font is used to indicate the best accuracy for each emotion, while underlining is used to indicate suboptimal accuracy. The results are presented in Tables 2 and 3. In the context of a six-category emotion setting, the MTG-ERC model demonstrated an accuracy of 70.12% and a weighted F1 score of 70.23%. This represents a notable improvement over the previous baseline model, with an increase of 0.99% in accuracy and 1.32% in weighted F1 score compared with the GraphCFC model. Among them, the F1 score for the happy and depressed categories are the best, while the F1 score for the sad, natural, angry and excited categories are the second best. In the four-category emotion setting, the MTG-ERC model demonstrates superior performance compared with the baseline model, attaining an accuracy of 84.77% and a weighted F1 score of 84.51%. This represents an improvement of 1.48% in accuracy and 1.3% in weighted F1 score compared with the COGMEN model.

Table 2. Experimental results of each model on the IEMOCAP (6-way) dataset.

Model	Happy	Sad	Neutral	Angry	Excited	Frustrated	Acc	w-F1
DialogueRNN9	33.18	78.80	59.21	65.28	71.86	58.91	63.40	62.75
DialogueGCN17	42.75	80.88	63.54	64.19	63.08	61.21	65.25	64.18
MMGCN18	42.34	78.67	61.73	69.00	74.33	62.32	65.43	66.22
ICON22	32.80	74.41	60.69	68.23	68.41	66.28	64.02	63.54
COGMEN23	51.90	81.70	68.60	66.00	75.30	58.20	68.20	67.60
GraphCFC19	43.23	84.45	64.13	71.37	78.19	62.74	69.13	68.91
MTG-ERC	55.59	82.09	65.62	69.62	75.56	66.64	70.12	70.23

Table 3. Experimental results of each model on the IEMOCAP (4-way) dataset.

Model	Acc	w-F1
BC-LSTM24	75.20	75.13
CHFusion25	76.59	76.80
COGMEN23	83.29	83.21
MTG-ERC	84.77	84.51

In this paper, we present an MTG-ERC model and experimentally compare it with the baseline method on the MELD dataset, which is used for multimodal emotion classification. The results are presented in Table 4. In the context of seven categories of emotions, the MTG-ERC model demonstrated an accuracy of 65.23% and a weighted F1 score of 65.32%. The highest accuracy was observed for the four categories of emotions: naturalness, fear, pleasure and disgust. It is noteworthy that the MTG-ERC model demonstrated superior performance in differentiating between similar emotions. Its F1 scores reached 18.31% and 15.30% for fear and disgust, respectively, showing strong discrimination ability.

Table 4. Experimental results of each model on the MELD dataset.

Model	Neutral	Surprise	Fear	Sadness	Joy	Disgust	Angry	Acc	w-F1
BC-LSTM24	73.80	47.70	5.40	25.10	51.30	5.20	38.40	57.50	55.90
DialogueRNN9	76.56	49.40	1.20	23.80	50.70	1.70	41.50	56.10	55.90
DialogueGCN17	-	-	-	-	-	-	-	57.97	58.10
MMGCN18	76.33	48.15	-	26.74	53.02	-	46.09	58.34	58.65
EmoCaps26	77.12	63.19	3.03	42.52	57.50	7.69	57.54	-	64.00
UniMSE15	-	-	-	-	-	-	-	65.09	65.51
GraphCFC19	75.48	48.30	1.90	26.33	51.58	3.30	46.41	61.42	62.08
MTG-ERC	78.91	59.72	18.31	40.23	64.10	15.30	54.96	65.25	65.32

Figure 3 depicts the confusion matrix representation of each emotion classification prediction value and label value for 6-way and 4-way in the IEMOCAP dataset. The confusion matrix serves as a visualization tool, facilitating a more intuitive understanding of the effect of emotion recognition. The colour depth in the matrix corresponds to the classification accuracy, with darker colour indicating higher accuracy.

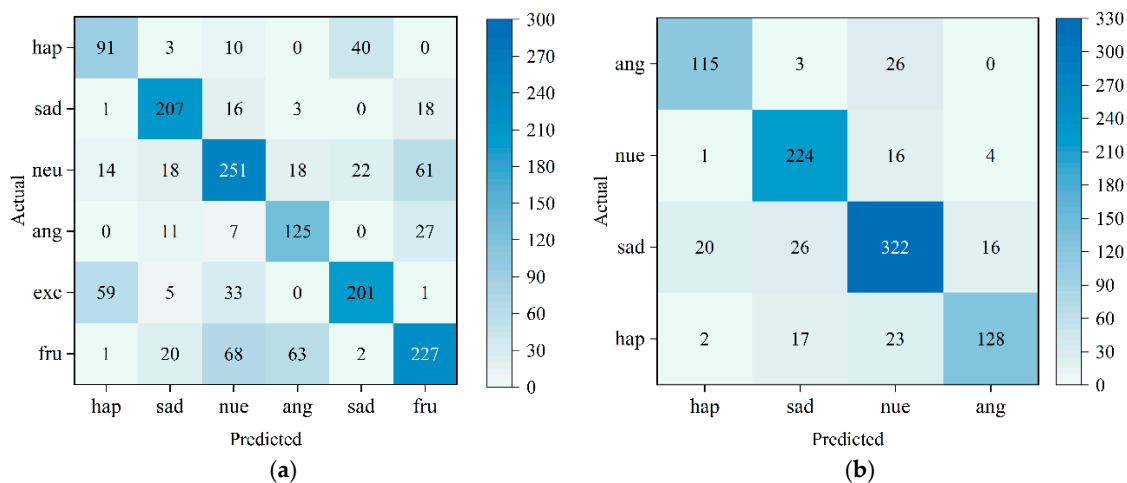


Figure 3. Confusion matrix of IEMOCAP (6-way) (a) and confusion matrix of IEMOCAP (6-way) (b).

4.4. Ablation Study

In order to evaluate the relative effectiveness of different modal tasks, the three modalities of text (T), audio (A) and vision (V) are combined and compared. The detailed performance results of each combination are presented in Table 5.

Table 5. Ablation experiments on each modality on IEMOCAP (6-way) and IEMOCAP (4-way) datasets.

Modality Settings	IEMOCAP (6-Way)		IEMOCAP (4-Way)	
	Acc	w-F1	Acc	w-F1
T	66.17	66.19	82.08	81.99
A	57.81	56.07	63.31	61.17
V	42.04	41.78	46.34	46.36
T + A	68.22	68.35	83.35	83.32
V + T	66.42	66.44	81.55	81.54
A + V	59.88	60.18	65.14	63.70
MTG-ERC	70.13	70.23	84.77	84.51

As demonstrated in Table 5, the accuracy and weighted F1 value of the multimodal emotion recognition task are superior to those of the unimodal task, thereby substantiating the significance of multimodal emotion recognition. Furthermore, the experimental results demonstrate that the unimodal tasks of text and audio contribute more to emotion recognition than the visual unimodal tasks.

To assess the efficacy of the multi-level Transformer fusion module and the directed multi-lateral relationship graph fusion module in multimodal emotion recognition, ablation comparison experiments were conducted on the components of these two modules.

As demonstrated in Table 6, the inter-transformer module, intra-transformer module and edge relationship module have been shown to significantly enhance the accuracy of the ERC task. The removal of a module resulted in a notable decline in model performance, with a reduction of 2.15% in accuracy and 2.11% in F1 score, respectively, on the IEMOCAP (6-way) dataset. This suggests that the module in the directed multilateral relationship graph fusion module plays a pivotal role in multi-modal conversational emotion recognition.

Table 6. Ablation experiments of each module on IEMOCAP (6-way) and IEMOCAP (4-way) datasets.

	IEMOCAP (6-Way)		IEMOCAP (4-Way)	
	Acc	w-F1	Acc	w-F1
w/o inter transformer	69.32	69.47	83.13	82.97
w/o intra transformer	68.44	68.63	82.38	82.16
w/o R_{inter}	68.35	68.52	82.02	81.97
w/o R_{intra}	67.98	68.12	82.43	82.37
MTG-ERC	70.13	70.23	84.77	84.51

The analysis in Table 5 indicates that the visual modality contributes less to the final performance compared to the textual and audio modalities. This observation can be attributed to several factors:

Inherent Information Density in Conversational Settings: In dyadic or multi-party conversations (as in IEMOCAP and MELD), the primary carriers of emotional information are often the linguistic content (text) and the paralinguistic cues (audio, e.g., tone, pitch). Visual cues, such as facial expressions, while valuable, can be more subtle, transient, or even intentionally suppressed in certain social contexts. Consequently, their standalone discriminative power for emotion classification is inherently lower than that of the other two modalities.

Challenges in Visual Data Quality: The visual stream is particularly susceptible to practical issues such as varying lighting conditions, partial occlusions, non-frontal face orientations, and lower effective resolution in video data. These factors introduce noise

and sparsity into the visual feature representations, making them less reliable than the relatively clean and stable features extracted from text and audio.

Limitations of Feature Representation: Our model relies on standard visual feature extractors (e.g., OpenFace). While these tools provide robust low-level facial action unit descriptors, they may not fully capture the high-level, dynamic, and complex temporal patterns of facial expressions that are crucial for distinguishing between certain fine-grained emotions (e.g., frustration vs. sadness). This represents a general challenge in the field rather than a specific shortcoming of our fusion architecture.

This discussion underscores that the lower contribution of the visual modality is likely a reflection of the dataset characteristics and inherent challenges in visual emotion recognition. Therefore, future work could focus on developing more sophisticated and dynamic visual feature extractors to better harness the complementary information that the visual channel provides.

4.5. Confusion Matrix Analysis

To gain deeper insights into the model's limitations and error patterns, we conducted a thorough analysis of the confusion matrices on the IEMOCAP dataset, as visualized in Figure 3. This analysis reveals which specific emotion pairs are most frequently confused by the MTG-ERC model.

In the 6-way classification setting (Figure 3a), we observe several consistent confusion patterns:

Happiness vs. Excitement: The most prominent confusion occurs between happy and excited. A significant number of excited utterances are misclassified as happy. This is likely because both emotions share similar high-arousal acoustic properties (e.g., increased pitch and speech rate) and often occur in positive contexts, making them challenging to distinguish, especially for audio-centric models.

Sadness vs. Frustration: There is a notable bidirectional confusion between sad and frustrated. Both are negative, low-valence emotions. The model sometimes struggles to capture the subtle differences, possibly because frustration can be a manifestation or a precursor to sadness in conversational contexts, and their textual expressions can be similar.

Anger vs. Frustration: The model also confuses angry with frustrated. These are both high-arousal, negative emotions. The distinction is often nuanced and context-dependent, relying on the intensity of the emotion, which the model may not always perfectly discern.

In the 4-way setting (Figure 3b), which consolidates some of the finer categories, the confusion between happy and the removed excited category is naturally absorbed, leading to a generally cleaner matrix. However, the challenge of distinguishing between the core negative emotions (sad, angry, frustrated consolidated) persists, indicating an inherent difficulty in fine-grained negative emotion classification.

Implications and Limitations: These findings highlight the primary limitation of our current model: its difficulty in disambiguating emotion pairs that are either acoustically proximate (e.g., happy/excited) or semantically and contextually similar in their negative valence (e.g., sad/frustrated/angry). This suggests that future work could benefit from designing more nuanced feature extractors capable of capturing subtle acoustic differences or incorporating external commonsense knowledge to better understand the contextual cues that differentiate these complex emotions.

5. Conclusions

The field of sentiment analysis is currently witnessing a growing interest in multimodal conversational emotion recognition, as the information derived from different modalities can reinforce each other, thereby enhancing the robustness of the system. This

paper presents a unified framework for the problem of conversational emotion recognition, designated MTG-ERC. This model is designed to capture the emotional information conveyed within and between modalities, as well as the emotional characteristics of global and local contexts, through the utilization of a multi-level Transformer fusion module and a directed multi-relational graph fusion module (for local, speaker-dependent conversational structure). Experimental results on the IEMOCAP and MELD datasets show consistent improvements (approximately 1% absolute in both accuracy and weighted F1) over a strong graph-based baseline. We further include ablation studies to isolate the contribution of each branch and to analyze robustness.

Author Contributions: Formal analysis, L.Y.; Data curation, C.W.; Writing—original draft, T.Y.; Writing—review & editing, H.J.; Project administration, X.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The original contributions presented in this study are included in the article. Further inquiries can be directed to the corresponding author.

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

References

1. Fu, Y.; Yuan, S.; Zhang, C.; Cao, J. Emotion Recognition in Conversations: A Survey Focusing on Context, Speaker Dependencies, and Fusion Methods. *Electronics* **2023**, *12*, 4714. [\[CrossRef\]](#)
2. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*. [\[CrossRef\]](#)
3. Busso, C.; Bulut, M.; Lee, C.-C.; Kazemzadeh, A.; Mower, E.; Kim, S.; Chang, J.N.; Lee, S.; Narayanan, S.S. IEMOCAP: Interactive emotional dyadic motion capture database. *Lang. Resour. Eval.* **2008**, *42*, 335–359. [\[CrossRef\]](#)
4. Poria, S.; Hazarika, D.; Majumder, N.; Naik, G.; Cambria, E.; Mihalcea, R. MELD: A Multimodal Multi-Party Dataset for Emotion Recognition in Conversations. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, 28 July–2 August 2019; pp. 527–536.
5. Zadeh, A.; Chen, M.; Poria, S.; Cambria, E.; Morency, L.P. Tensor Fusion Network for Multimodal Sentiment Analysis. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Copenhagen, Denmark, 9–11 September 2017; pp. 1103–1114.
6. Liu, Z.; Shen, Y.; Lakshminarasimhan, V.B.; Liang, P.P.; Zadeh, A.B.; Morency, L.P. Efficient Low-rank Multimodal Fusion With Modality-Specific Factors. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, Melbourne, Australia, 15–20 July 2018; pp. 2247–2256.
7. Tsai, Y.H.H.; Bai, S.; Liang, P.P.; Kolter, J.Z.; Morency, L.P.; Salakhutdinov, R. Multimodal Transformer for Unaligned Multimodal Language Sequences. In Proceedings of the Conference Association for Computational Linguistics, Florence, Italy, 28 July–2 August 2019; pp. 6558–6569.
8. Sahay, S.; Okur, E.; Kumar, S.H.; Nachman, L. Low Rank Fusion based Transformers for Multimodal Sequences. In *Second Grand-Challenge and Workshop on Multimodal Language (Challenge-HML)*; ACL: Seattle, WA, USA, 2020; pp. 29–34.
9. Majumder, N.; Poria, S.; Hazarika, D.; Mihalcea, R.; Gelbukh, A.; Cambria, E. Dialoguernn: An attentive rnn for emotion detection in conversations. In Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, 27 January–1 February 2019; Volume 33, pp. 6818–6825.
10. Ghosal, D.; Majumder, N.; Gelbukh, A.; Mihalcea, R.; Poria, S. COSMIC: COMmonSense knowledge for eMotion Identification in Conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2020*; ACL: Seattle, WA, USA, 2020; pp. 2470–2481.
11. Hu, D.; Wei, L.; Huai, X. DialogueCRN: Contextual Reasoning Networks for Emotion Recognition in Conversations. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, Online, 1–6 August 2021; pp. 7042–7052.

12. Shen, W.; Chen, J.; Quan, X.; Xie, Z. Dialogxl: All-in-one xlnet for multi-party conversation emotion recognition. In Proceedings of the AAAI Conference on Artificial Intelligence, Online, 2–9 February 2021; Volume 35, pp. 13789–13797.
13. Kim, T.; Vossen, P. Emoberta: Speaker-aware emotion recognition in conversation with roberta. *arXiv* **2021**, arXiv:2108.12009.
14. Li, S.; Yan, H.; Qiu, X. Contrast and generation make bart a good dialogue emotion recognizer. In Proceedings of the AAAI Conference on Artificial Intelligence, Online, 22 February–1 March 2022; Volume 36, pp. 11002–11010.
15. Hu, G.; Lin, T.E.; Zhao, Y.; Lu, G.; Wu, Y.; Li, Y. UniMSE: Towards Unified Multimodal Sentiment Analysis and Emotion Recognition. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, Abu Dhabi, United Arab Emirates, 7–11 December 2022; pp. 7837–7851.
16. Ghosal, D.; Majumder, N.; Poria, S.; Chhaya, N.; Gelbukh, A. DialogueGCN: A Graph Convolutional Neural Network for Emotion Recognition in Conversation. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Hong Kong, China, 3–7 November 2019; pp. 154–164.
17. Ishiwatari, T.; Yasuda, Y.; Miyazaki, T.; Goto, J. Relation-aware graph attention networks with relational position encodings for emotion recognition in conversations. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Online, 16–20 November 2020; pp. 7360–7370.
18. Hu, J.; Liu, Y.; Zhao, J.; Jin, Q. MMGCN: Multimodal Fusion via Deep Graph Convolution Network for Emotion Recognition in Conversation. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, Online, 1–6 August 2021; pp. 5666–5675.
19. Li, J.; Wang, X.; Lv, G.; Zeng, Z. GraphCFC: A Directed Graph Based Cross-Modal Feature Complementation Approach for Multimodal Conversational Emotion Recognition. *IEEE Trans. Multimed.* **2023**, *26*, 77–89. [[CrossRef](#)]
20. Schlichtkrull, M.; Kipf, T.N.; Bloem, P.; Van Den Berg, R.; Titov, I.; Welling, M. Modeling relational data with graph convolutional networks. In *The Semantic Web, Proceedings of the 15th International Conference, ESWC 2018, Heraklion, Greece, 3–7 June 2018*; Springer International Publishing: Berlin/Heidelberg, Germany, 2018; pp. 593–607.
21. Yun, S.; Jeong, M.; Kim, R.; Kang, J.; Kim, H.J. Graph transformer networks. *Adv. Neural Inf. Process. Syst.* **2019**, *32*. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.