Hybrid Relational Graphs with Sentiment-laden Semantic Alignment for Multimodal Emotion Recognition in Conversation

Hongru Ji[1] , **Xianghua Li**[1] , **Mingxin Li**[1] , **Meng Zhao**[2] and **Chao Gao**[1*]

[1]Northwestern Polytechnical University, Shaanxi, China
[2]Henan University of Technology, Zhengzhou, China
{jihongru,mingxinli}@mail.nwpu.edu.cn, zm@haut.edu.cn, {li_xianghua,cgao}@nwpu.edu.cn

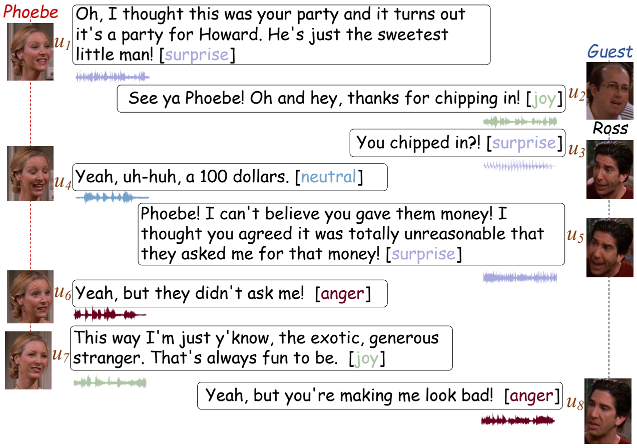
Abstract

Multimodal Emotion Recognition in Conversation (MERC) focuses on detecting the emotions expressed by speakers in each utterance. Recent research has increasingly leveraged graph-based models to capture interactive relationships in conversations, enhancing the ability to extract emotional cues. However, existing methods primarily focus on explicit utterance-level relationships, neglecting both the implicit connections within individual modality and the differences in implicit relationships across modalities. Moreover, these methods often overlook the role of sentimental features in conversation history in cross-modal semantic alignment. To address these issues, we propose a novel model that employs modality-adaptive hybrid relational graphs to enrich the dialogue graph by inferring implicit relationships between nodes within each modality. Furthermore, we introduce historical sentiment through a progressive strategy that utilizes contrastive learning to refine cross-modal semantic alignment. Experimental results demonstrate the superior performance of our approach over state-of-the-art methods on the IEMOCAP and MELD datasets. Our code is available at https://github.com/cgao-comp/HRG-SSA.

1 Introduction

Emotions significantly shape human interactions, decisions, and overall well-being [Baltrušaitis *et al.*, 2019]. With the growing expectation for machines to perceive, understand, and express emotions like humans, detecting emotions from conversations using multimodal information has become increasingly crucial [Geetha *et al.*, 2024].

Multimodal emotion recognition in conversation (MERC) aims to utilize text, acoustic, and visual modalities within a conversational context to accurately identify the emotion expressed by the speaker in each utterance. Typically, MERC is treated as a classification task, where researchers often start with efficient multimodal information processing and fusion, followed by applying a classifier for emotion detection, as demonstrated in [Li *et al.*, 2023a; Hu *et al.*, 2024]. Recently, some studies have explored the use of text generation for emotion recognition. For example, [Hu *et al.*, 2022b] unified sentiment analysis and emotion recognition by designing text labels and employing Transformer decoders [Vaswani *et al.*, 2017] to progressively generate both the results and confidence scores. Based on this, [Li *et al.*, 2023b] introduces UniSA, a unified multimodal generative framework designed to effectively handle various emotion-related detection tasks.



Oh, I thought this was your party and it turns out it's a party for Howard. He's just the sweetest little man! [surprise]

See ya Phoebe! Oh and hey, thanks for chipping in! [joy]

You chipped in?! [surprise]

Yeah, uh-huh, a 100 dollars. [neutral]

Phoebe! I can't believe you gave them money! I thought you agreed it was totally unreasonable that they asked me for that money! [surprise]

Yeah, but they didn't ask me! [anger]

This way I'm just y'know, the exotic, generous stranger. That's always fun to be. [joy]

Yeah, but you're making me look bad! [anger]

Figure 1: An example of utterances in a conversation, including text, acoustic, and visual modalities. The emotional labels of utterances are highlighted in different colors.

Alternatively, to address the complex relationships and dependencies among diverse modalities, graph-based methods utilize graph structures to capture intricate interactions and offer a comprehensive understanding of sentiment. Some standard graphs define specific rules to explicitly identify the relationships between nodes. [Li *et al.*, 2024] establishes connections between nodes from different modalities; for instance, a textual node is linked to all visual nodes. [Nguyen *et al.*, 2023] connects nodes within a specific window of the same modality, and nodes representing different modalities of the same utterance. Additionally, [Yi *et al.*, 2024] replaces the aforementioned relationships with two types of corresponding hyperedges, which can connect any number of nodes, thereby naturally encoding relationships of higher arity.

However, they have limitations in graph structure and over-

[*]Corresponding Author

look certain aspects of sentiment encoding: (1) **Sentiment features in the conversation history are ignored.** The classification task generates discrete labels that do not inherently convey emotional content. For instance, when ”0” is assigned to represent happiness, no emotion-related information can be derived from the label ”0” in isolation. Consequently, the emotional features detected in the conversation history are not effectively utilized in predicting the emotions of subsequent utterances. However, incorporating historical emotions can provide benefits(e.g. Utterances 6 & 8 in Figure 1). (2) **Implicit relationships between utterances in the conversation graph are overlooked.** Blindly connecting all nodes within the same conversation may introduce noise by linking semantically unrelated nodes (Utterances 2 & 8, etc.), thereby influencing emotion recognition accuracy. Conversely, establishing connections solely based on explicit relationships that match predefined rules may lead to the neglect of implicit relationships (Utterances 4 & 8, etc.), which are equally crucial for comprehensive emotional analysis. (3) **Variations in implicit relationships across different modalities are disregarded.** In fact, there are distinct differences in the implicit relationships between nodes across various modalities. For example, two speakers may appear simultaneously in a scene and interact physically, but not verbally. This creates a relationship within the visual modality, while no such relationship exists in the textual or auditory modalities. Applying the relational structure of one modality uniformly across others may result in the introduction of spurious connections or the loss of crucial contextual information. Existing research has yet to provide a solution to these discrepancies.

To address the challenges outlined earlier, we propose a novel encoder-decoder framework, HRG-SSA, which combines Hybrid Relational Graphs(HRG) with a Sentiment-laden Semantic Alignment strategy(SSA). This approach treats MERC as a text generation task, thereby endowing the emotional labels in the conversation history with genuine textual emotional semantics. SSA employs a two-stage contrastive learning process. Initially, it aligns the various modal features of the conversation history with emotions, and subsequently, it aligns the semantics of each sentiment-laden modal feature. This progressive alignment enhances the model's ability to fuse multimodal information during the encoding process. Inspired by [Zhao *et al.*, 2022], HRG employs a modality-adaptive connection prediction module that infers implicit relationships between nodes within a modality based on their semantic similarity. It then integrates these inferred relationships with explicit ones to construct a more coherent and comprehensive conversation graph structure, thereby enhancing the interaction and fusion of features within each modality. Ultimately, a decoder incorporating dynamic fusion layers is employed to effectively integrate multimodal information, enabling the generation of the emotional polarity for the current utterance. The main contributions of this paper can be summarized as follows:

- We propose an SSA strategy that uses progressive contrastive learning to integrate sentiment cues into contextual references, facilitating multimodal semantic alignment. Additionally, a novel modality-adaptive HRG infers implicit connections within modalities, completing the dialogue graph and improving modality fusion.
- We propose an end-to-end generation framework HRG-SSA that incorporates HRG and SSA into the encoding phase to enhance multimodal information representation. The decoder subsequently employs dynamic fusion layers to seamlessly integrate multimodal inputs and generate the speaker's emotional tendency.
- Experiments on the IEMOCAP and MELD datasets demonstrate that our method outperforms state-of-the-art approaches in accuracy and weighted F1 score, highlighting its effectiveness in MERC.

2 Related Work

2.1 Transformer-Based Models for MERC

Drawing inspiration from Transformer architectures, most of recent studies focus on designing cross-modal attention mechanisms to effectively fuse various modalities [Rahman *et al.*, 2020; Huang *et al.*, 2023; Sun *et al.*, 2023]. [Hubert *et al.*, 2019] categorized the challenges in multimodal data modeling into two types: misalignment within modalities and across modal feature spaces. They proposed the end-to-end Multimodal Transformer (MulT) to address these issues and showed the effectiveness of the attention mechanism in modal fusion. CENet enhances each word's representation by incorporating long-range emotional cues implicit in the visual and auditory modalities [Wang *et al.*, 2022]. SDT employs a Transformer-based model with intra-modal and intermodal attention mechanisms to capture interactions and dynamically learn modality weights through hierarchical gating [Ma *et al.*, 2023]. [Maji *et al.*, 2023] designed a cross-modal transformer block to capture interactions and temporal dependencies between audio and text, while using self-attention to prioritize key sentiment information from the fused features. Recent studies have explored text generation for emotion recognition using transformer decoders. They unified the emotion-related classification tasks into an end-to-end framework by unifying their label formats [Hu *et al.*, 2022b; Li *et al.*, 2023b]. Although these methods have made progress in MERC, none have recognized the potential of emotional labels in dialogue history to enhance emotion prediction in subsequent interactions.

2.2 Graph-Based Models for MERC

Sequential modeling has an upper limit in extracting features from conversation context, prompting some researches to explore the graph structure within dialogues. These approaches typically treat each modality of a utterance as a node and establish connections based on meaningful relationships [Hu *et al.*, 2024]. Early studies relied on the adjacency relationships inherent in the conversational structure to establish edges. For example, a common approach is to connect the nodes within the same modality for a given conversation and link nodes across different modalities within the same utterance [Hu *et al.*, 2021a; Nguyen *et al.*, 2023]. Moreover, some studies exploit speaker-specific relationships to establish connections between the corresponding nodes [Lian *et al.*, 2023; Lee and Choi, 2021]. Furthermore, some researchers have

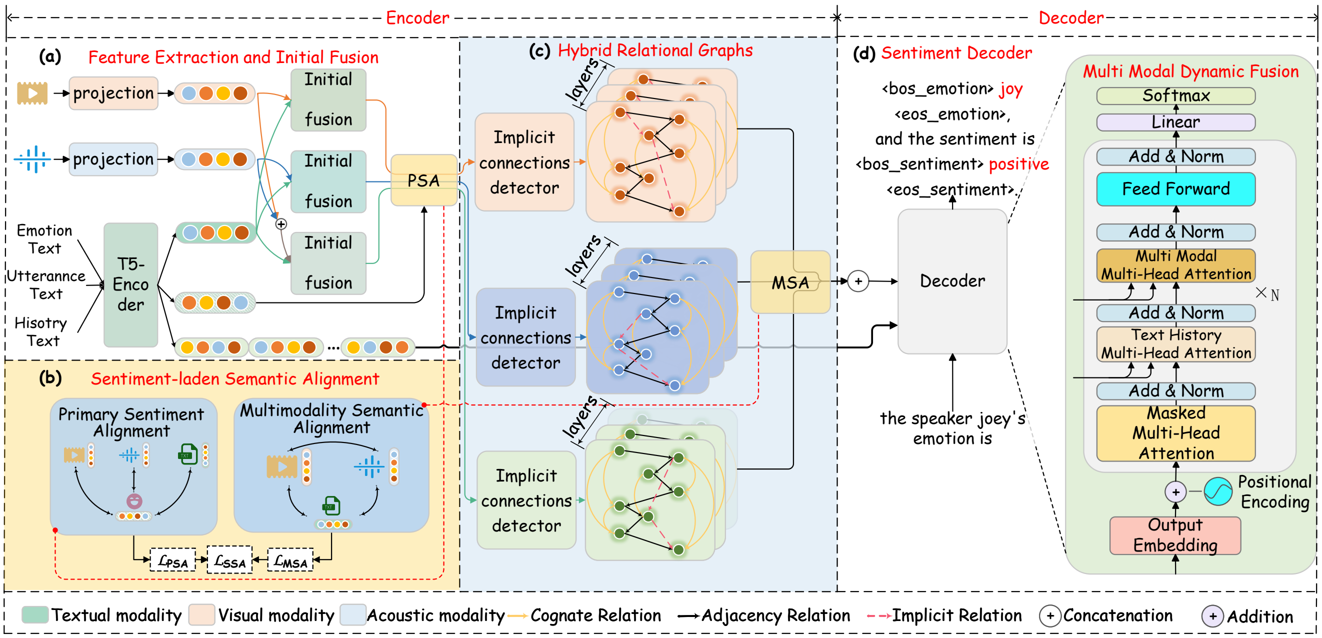

Encoder
Decoder
(a) Feature Extraction and Initial Fusion
projection
projection
Initial fusion
Initial fusion
Initial fusion
Emotion Text
Utterannce Text
Hisotry Text
T5-Enco der
PSA
(c) Hybrid Relational Graphs
Implicit connections detector
Implicit connections detector
Implicit connections detector
layers
layers
layers
MSA
(b) Sentiment-laden Semantic Alignment
Primary Sentiment Alignment
Multimodality Semantic Alignment
$\mathcal{L}_{PSA}$
$\mathcal{L}_{SSA}$
$\mathcal{L}_{MSA}$
(d) Sentiment Decoder
<bos_emotion> joy <eos_emotion>, and the sentiment is <bos_sentiment> positive <eos_sentiment>.
Decoder
the speaker joey's emotion is
Multi Modal Dynamic Fusion
Softmax
Linear
Add & Norm
Feed Forward
Add & Norm
Multi Modal Multi-Head Attention
×N
Add & Norm
Text History Multi-Head Attention
Add & Norm
Masked Multi-Head Attention
Positional Encoding
Output Embedding
Textual modality
Visual modality
Acoustic modality
Cognate Relation
Adjacency Relation
Implicit Relation
Concatenation
Addition

Figure 2: Illustration of HRG-SSA framework. In the encoder, text, acoustic, and visual features are represented by green, blue, and orange, respectively. Within the graphs structure, black solid lines denote adjacency relationships, yellow solid lines connect nodes from the same speaker, and red dotted lines denote implicit relationships within each modality. In the decoder, N represents the number of stacked blocks.

exploratively leveraged external knowledge and rules to extract relationships within the context. For example, M3GAT employs the spaCy toolkit to construct a dependency tree of text utterances, leveraging the syntactic dependencies between words to build edges [Zhang *et al.*, 2023a]. DER-GCN constructs a multi-relational sentiment interaction graph that incorporates relationships between speakers and heterogeneous elements derived from events [Ai *et al.*, 2024]. MKE-IGN integrates textual and visual common-sense knowledge into the edge representation to improve the modeling of conversation graph relations by external knowledge [Tu *et al.*, 2024a]. The introduction of hyperedges effectively shortens the distance between interactions across different modalities, making the hypergraph an innovative approach for modeling complex relationships in MERC [Chen *et al.*, 2023; Yi *et al.*, 2024]. Exploratively, [Pei *et al.*, 2024] proposed a multi-track graph convolutional network that separates message passing by category semantics to prevent heterophilic mixing and mitigate oversmoothing and oversquashing in deep GNNs. However, these methods mainly focus on explicit utterance-level relationships, overlooking both the implicit connections within each modality and the variations in implicit relationships across modalities.

3 Methodology

In this section, we provide a detailed overview of each module in the proposed HRG-SSA, as illustrated in Figure 2.

3.1 Problem Formulation

A conversation consists of a sequence of utterances: $C = \{u_1, u_2, ..., u_N\}$. N is the number of utterances. Each utterance u_i consists of three modalities: text, audio, and vision, along with a corresponding speaker: $u_i = \{u_i^t, u_i^a, u_i^v; S_i\}$. $u_i^t \in \mathbb{R}^{d_t}$, $u_i^a \in \mathbb{R}^{d_a}$ and $u_i^v \in \mathbb{R}^{d_v}$, d_t, d_a, d_v denote represent the dimensionalities of the text, audio, and visual modalities, respectively. The goal of MERC is to predict the emotion y_i expressed by the speaker S_i in each utterance u_i, based on the conversation history: $C_h = \{u_1, u_2, ..., u_{i-1}\}$.

3.2 Feature Extraction and Encoding

Following [Chen *et al.*, 2023], we employ DenseNet [Huang *et al.*, 2017] or 3D-CNN [Yang *et al.*, 2019] to extract features from the visual modality, and utilize OpenSmile toolkit [Eyben *et al.*, 2010] to extract features from the audio modality. For the text modality, we uniformly utilize the pre-trained T5-base [Raffel *et al.*, 2020] encoder for feature extraction.

Unimodal Encoder: To ensure consistency in the dimensionality of the representations across modalities, we employ two projection layers to perform preliminary spatial alignment for the audio and visual features, as detailed below:

$$\begin{aligned} h_i^a &= W_a u_i^a + b_i^a \\ h_i^v &= W_v u_i^v + b_i^v \end{aligned} \tag{1}$$

where u_i^η represents the extracted feature, resulting in fixed-size representations $h_i^\eta \in \mathbb{R}^{d_h}, \eta \in \{a, v\}$; W_η and b^η are learnable parameters. It is worth noting that these two modalities are only encoded at the utterance level.

The text modality includes both utterance-level and context-level encoding, which are utilized for graph initialization and the cross-attention module of the decoder, respectively. We integrate speaker identity information by adding a "speaker:" prefix to each utterance, followed by an end-of-sequence token "</s>" as a suffix. For instance, the Utterance 3 in Figure 1 appears as "Ross: You chipped in?!

</s>". We use the representation of the token "</s>" as the utterance-level encoding, while retaining the context-level representations:

$$h_i^t = \text{T5Encoder}(u_i^t)[indexof(\text{"</s>"})] \tag{2}$$

$$h_C^t = \text{T5Encoder}(C_h^t) \tag{3}$$

where $C_h^t = \{u_1^t, u_2^t, ..., u_{i-1}^t\}$, $indexof(\text{"</s>"})$ represents the index of "</s>".

Sentiment Encoding: To effectively leverage the sentiments within the conversation history, we structured the sentimental labels of each utterance into a standardized textual format. The specific format is as follows: "Ross: surprise & negative </s>", and we use s_i to denote that of utterance i. This structured representation ensures consistent integration of emotional context throughout the dialogue. Encoding it with the T5 encoder yields a representation rich in emotional semantics, enabling a more nuanced understanding of the dialogue's sentiment:

$$h_i^s = \text{T5Encoder}(s_i) \tag{4}$$

3.3 Initial Fusion and Sentiment Alignment

After encoding, we employ the cross-attention (CS-Attention for short) [Vaswani *et al.*, 2017] mechanism to facilitate the initial fusion between modalities at the utterance level. For the text modality, we use h_i^t as the query, while the concatenation of h_i^a and h_i^v serves as both the key and value for attentional interactions.

$$h^{av} = [h^a \parallel h^v] \tag{5}$$

$$\begin{aligned} H^t &= \text{CS-Attention}(h^t, h^{av}, h^{av}) \\ &= \text{Softmax}(\frac{(h^t W_t^Q)(h^{av} W_t^K)^T}{\sqrt{d}})(h^{av} W_t^V) \end{aligned} \tag{6}$$

where $\parallel$ denotes the concatenate operation; $\cdot^T$ represents the transpose operation of a matrix; W_t^Q, W_t^K, W_t^V are learnable parameter matrices; d is the dimension of hidden states.

In order to improve fusion efficiency and align with the conclusion that text plays a dominant role in multimodal affective computing [Zhang *et al.*, 2023a; Zhang *et al.*, 2023b], we use h_i^t as the key and value for acoustic and visual modalities. And we utilize h_i^a and h_i^v as the queries for attentional interactions, respectively.

$$\begin{aligned} H^\eta &= \text{CS-Attention}(h^\eta, h^t, h^t) \\ &= \text{Softmax}(\frac{(h^\eta W_\eta^Q)(h^t W_\eta^K)^T}{\sqrt{d}})(h^t W_\eta^V) \end{aligned} \tag{7}$$

where $\eta \in \{a, v\}$; $W_\eta^Q, W_\eta^K, W_\eta^V$ are learnable parameter matrices.

Primary Sentiment Alignment(PSA): In this module, we use historical sentiment to align contexts with their corresponding sentiment expressions. Multiple sentiments are often present in a conversation, and aligning the semantics of an utterance with its true emotional state, while distancing it from other emotions, enhances the model's attention to sentimental cues, thereby improving subsequent emotion prediction. In prior work, we encoded both the speaker's utterances and their associated emotions to obtain h_i^s. In the current module, we treat each modal representation $H_i^\eta, \eta \in \{t, a, v\}$ and its corresponding h_i^s as a positive pair, while pairing h_i^η with the representations of other $h_j^s, j \neq i$ as negative pairs. This contrastive learning strategy facilitates semantic alignment across modalities. Suppose there are n utterances in the dialogue, and the comparative loss is as follows:

$$\mathcal{L}_{PSA} = -\sum_{i=1}^{n-1} \sum_{\eta \in \{t,a,v\}} \log \frac{e^{sim(H_i^\eta, h_i^s)/\tau}}{\sum_{j \neq i}^{n-1} e^{sim(H_i^\eta, h_j^s)/\tau}} \tag{8}$$

where the target utterance is not involved; $sim(H_i^\eta, h_i^s)$ denotes the cosine similarity between two vectors; $\log(\cdot)$ denotes the logarithmic function; and τ is a temperature parameter. By minimizing this loss, modality features are effectively aligned with the correct sentiment representation space.

3.4 Hybrid Relational Graphs

Graph Structure: We represent successive conversation contexts using a hybrid relational directed graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where each node $v \in \mathcal{V}$ represents a unimodal utterance and the set of edges $\mathcal{E}$ includes both explicit and implicit relationships between the nodes. All nodes are initialized with their corresponding hidden states H_i^η. And, the directed graph structure ensures that subsequent nodes cannot be used for emotion prediction of preceding nodes, preserving the temporal causality of the conversation.

Explicit relations remain consistent across modalities; we separately link neighboring utterances (including the utterance itself) and those from the same speaker. This approach preserves the temporal relationships underlying the pairwise connections and enables the tracking of each speaker's emotional evolution. **Implicit connections(IC)** capture the hidden pathways of information propagation within each modality by employing a detector that evaluates the correlations between nodes. These connections complement explicit relationships and result in modality-specific edge structures.

Specifically, based on the hidden states of nodes in each modality, the implicit unobserved relationships of modality adaptation are computed. A single-layer feedforward neural network, along with a mask matrix, is employed to calculate the scores s_{ij}^η between node i and j, thereby inferring those implicit relationships.

$$s_{ij}^\eta = \text{LeakyRelu}\left(mask + W_\eta^T \left[H_i^\eta \parallel H_j^\eta\right]\right) \tag{9}$$

where $\eta \in \{t, a, v\}$; W_η^T is a learnable parameter matrix; LeakyReLU is a nonlinearity activation function; and $mask \in \mathbb{R}^{n*n}$ is a square matrix with $-inf$ in the upper triangle and zeros in the lower triangle, thus preventing backward dependency. Then, a softmax function is applied to calculate the attention values α_{ij}^η for each node pair.

$$\alpha_{ij}^\eta = \frac{\exp(s_{ij}^\eta)}{\sum_{k=0}^{n} \exp(s_{kj}^\eta)} \tag{10}$$

where $\exp(\cdot)$ represents the exponential function. If the value exceeds the average score of node j predecessor, an implicit connection is established between node i and j $(i<j)$:

$$A_{i,j} = \begin{cases} 1, & \text{if } \alpha_{ij}^\eta > \frac{1}{j} \\ 0, & \text{otherwise} \end{cases} \tag{11}$$

where A represents the adjacency matrix.

Graph Update: The update of node representations within the graph structure follows the graph attention network [Veličković *et al.*, 2018] framework, comprising two steps: directed edge weight calculation and information aggregation. First, we take each node representation H_i^η as input and apply the self attention mechanism [Vaswani *et al.*, 2017] to compute the attention distribution between a specific node and its previously connected nodes.

$$\begin{aligned}\hat{s}_{ij}^{\eta} &= \frac{(W_e^\eta H_i^\eta W_e^Q)(W_e^\eta H_j^\eta W_e^K)^T}{\sqrt{d}} \\ \hat{\alpha}_{ij}^{\eta} &= \frac{\exp(\hat{s}_{ij}^{\eta})}{\sum_{k\in\mathcal{N}_j}\exp(\hat{s}_{kj}^{\eta})}\end{aligned} \tag{12}$$

where W_e is a learnable linear transformation, aimed at providing sufficient expressive power to map the input features into higher-level representations; W_e^Q and W_e^K are learnable parameter matrices for attention mechanism; $\mathcal{N}_j$ denotes the neighborhood of node j in the graph.

Based on the attention weights computed above, we update each node by aggregating the features of its neighboring nodes. Additionally, we stack L layers of graph updates to facilitate information exchange across multi-hop nodes:

$$H_i^{\eta^l} = \sigma\left(\sum_{j\in\mathcal{N}_i}\hat{\alpha}_{ij}^{\eta^l} W_e^{\eta^l} H_j^{\eta^{l-1}}\right) \tag{13}$$

where η represents the modality; l denotes the index of the layer; $\sigma(\cdot)$ denotes the sigmoid function. We use the output of the final layer as the encoded representation of the hybrid relationship graph, denoted as $H_i^{\eta^L}$.

Multimodality Semantic Alignment(MSA): Through the previous operations, the utterance node representations have been enriched with emotional context, and sufficient information fusion within the modality has been achieved. In this module, we will once again employ contrastive learning to perform cross-modal semantic alignment. The three modalities from the same modality form positive pairs pairwise, while negative pairs are formed between different utterance nodes. The comparative loss is as follows:

$$\mathcal{L}_{MSA} = -\sum_{i=1}^{n}\sum_{\eta\in\{t,a,v\}}\sum_{\xi\neq\eta}\log\frac{e^{sim(H_i^\eta,H_i^\xi)/\tau}}{\sum_{j\neq i}^{n} e^{sim(H_i^\eta,H_j^\xi)/\tau}} \tag{14}$$

where n denotes the number of nodes in the graph; ξ indicates two other modalities besides η. By minimizing this loss function, the semantics of multiple modalities for the same node are brought into closer alignment.

Up to this point, the previous PSA module and the current MSA module progressively refine the role of sentiment cues, further promoting cross-modal alignment.

3.5 Decoder

This framework employs text generation tasks for emotion classification. We organized the sentimental labels into the fluent text format: "In the last round of the above dialogue, the speaker Ross's emotion is <bos_emotion>surprise <eos_emotion>, and sentiment is <bos_sentiment>negative <eos_sentiment>.", where the special tokens with pointed brackets are placeholders for label extraction.

MultiModal Dynamic Fusion: Specifically, we utilize the pre-trained T5-base Decoder as the decoder and introduce several enhancements to optimize its performance. As mentioned previously, we have multiple inputs: h_C^t, H^{η^L}, and H_D(representing the hidden states of the decoder input). Since the classic transformer decoder can only process two inputs with a single cross-attention layer, we introduce a multimodal dynamic fusion structure by adding an additional cross-attention layer in series.

First, we use $\hat{H}_D$, obtained from the masked self-attention layer, as the query, while the conversation history representation h_C^t serves as the key and value, passing through the first cross-attention layer. This approach allows us to capture event interactions and sentiment cues from the conversation history:

$$\begin{aligned}H_{DC} &= \text{CS-Attention}(\hat{H}_D, h_C^t, h_C^t) \\ &= \text{Softmax}(\frac{(\hat{H}_D W_c^Q)(h_C^t W_c^K)^T}{\sqrt{d}})(h_C^t W_c^V)\end{aligned} \tag{15}$$

where W_c^Q, W_c^K and W_c^V are learnable matrices. Next, we use the concatenation of H^{η^L} as both the key and value, with H_{DC} as the query, to extract sentiment cues from the multimodal representation through the second cross-attention layer:

$$H^{tav} = [H^{t^L} \parallel H^{a^L} \parallel H^{v^L}] \tag{16}$$

$$\begin{aligned}H_{DCM} &= \text{CS-Attention}(H_{DC}, H^{tav}, H^{tav}) \\ &= \text{Softmax}(\frac{(H_{DC} W_m^Q)(H^{tav} W_m^K)^T}{\sqrt{d}})(H^{tav} W_m^V)\end{aligned} \tag{17}$$

where W_m^Q, W_m^K and W_m^V are learnable matrices. Once the output is obtained, it is mapped onto the vocabulary through a feedforward network (FFN), and the probability distribution is generated using softmax, consistent with the original Transformer decoder:

$$P^{voc}(y_t) = \text{Softmax}(W_s H_{DCM}) \tag{18}$$

where t denotes the t-th time step and W_s is learnable matrix. We adopt a cross-entropy loss to optimize this generation task:

$$\mathcal{L}_s = -\sum_{t=1}\log\left(P^{voc}(y_t)\right) \tag{19}$$

Training Objectives: The final training objective is a weighted fusion of the three loss functions described above:

$$\mathcal{L} = \mathcal{L}_s + \lambda(\mathcal{L}_{PSA} + \mathcal{L}_{MSA}) \tag{20}$$

where the hyperparameter weight λ controls the significance of the contrastive loss.

4 Experiments

4.1 Experimental Settings

Datasets: We evaluate the performance of our model on two popular datasets for MERC: IEMOCAP [Busso *et al.*, 2008]

Methods	IEMOCAP								MELD								
	Happy	Sad	Neutral	Angry	Excited	Frustrated	Acc	w-F1	Neutral	Surprise	Fear	Sadness	Joy	Disgust	Anger	Acc	w-F1
UniMSE♣	-	-	-	-	-	-	70.56	70.66	-	-	-	-	-	-	-	65.09	65.51
UniSA♣	-	-	-	-	-	-	70.56	71.77	-	-	-	-	-	-	-	67.85	66.71
DialogueGCN♢	51.57	80.48	57.69	53.95	72.81	57.33	63.22	62.89	75.97	46.05	-	19.60	51.20	-	40.83	58.62	56.36
MMGCN♢	45.14	77.16	64.36	68.82	74.71	61.40	66.36	66.26	76.33	48.15	-	26.74	53.02	-	46.09	60.42	58.31
MMDFN♢	42.22	78.98	66.42	69.77	75.56	66.33	68.21	68.18	77.76	50.69	-	22.93	54.78	-	47.82	62.49	59.46
AdaIGN♣	53.04	81.47	71.26	65.87	76.34	67.79	-	70.74	79.75	**60.53**	-	**43.70**	64.54	-	56.15	-	66.79
DER-GCN♣	58.8	79.8	61.5	**72.1**	73.3	67.8	69.7	69.4	**80.6**	51.0	10.4	41.5	64.3	10.3	**57.4**	66.8	66.1
M^{3}Net♠	54.43	81.24	66.16	65.42	75.22	63.57	68.21	68.38	79.99	60.14	15.38	35.03	64.50	26.67	54.23	67.59	65.85
HAUCL♠	60.73	77.10	69.74	65.34	77.69	60.97	69.19	68.99	80.06	60.25	15.87	35.03	**64.68**	28.30	54.15	67.70	65.97
HRG-SSA-gen	66.47	83.16	73.94	68.31	82.43	**69.63**	74.37	74.63	80.58	58.99	13.56	41.26	63.53	26.37	54.15	67.97	66.30
HRG-SSA	**71.27**	**84.79**	**74.50**	70.59	**82.68**	68.56	**75.48**	**75.47**	80.46	59.85	**21.05**	41.95	63.40	**30.63**	55.62	**68.05**	**66.83**

♢, ♣, ♠ indicates that the results come from [Hu *et al.*, 2022a], original papers and our replication, respectively.

Table 1: Results on IEMOCAP and MELD. Bolded values indicate the best results, and underlined values denote the second-best.

and MELD [Poria *et al.*, 2019]. The former consists of videos of two-person conversations between 10 actors, while the latter is a multimodal dataset for multi-party conversations, taken from the Friends TV series. The distribution of both datasets is presented in Table 2. Note that since IEMOCAP does not provide labels for emotional stance, we assigned positive, negative, or neutral based on their emotional labels.

Dataset	Dialogues			Utterances			Classes
	train	val	test	train	val	test	
IEMOCAP	120	31		5810	1623		6
MELD	1039	114	280	9989	1109	2610	7

Table 2: Dataset statistics for IEMOCAP and MELD.

Implementation Details: All experiments were conducted on a single NVIDIA RTX A6000. We used a two-stage optimization strategy, first fine-tuning non-pretrained parameters (10 epochs) and then optimizing all parameters (10 epochs), with learning rates set to 5e-5 and 3e-5, respectively. Adam optimizer is used for training. The values of the remaining key hyperparameters are shown in the Table 3.

Dataset	Epoch	Batch size	λ	L	τ	Warmup ratio
IEMOCAP	20	8	0.001	1	1.0	0.2
MELD	20	8	0.001	3	1.0	0.2

Table 3: Main hyperparameters for HRG-SSA.

Baselines and Metrics: To evaluate the performance of our approach, we compare it with the following state-of-the-art methods: (1) Transformer-Based models: UniMSE [Hu *et al.*, 2022b] and UniSA [Li *et al.*, 2023b]; (2) Graph-Based models: DialogueGCN [Ghosal *et al.*, 2019], MMGCN [Hu *et al.*, 2021b], MMDFN [Hu *et al.*, 2022a], M^{3}Net [Chen *et al.*, 2023], AdaIGN [Tu *et al.*, 2024b], DER-GCN [Ai *et al.*, 2024], HAUCL [Yi *et al.*, 2024]. For details on the baseline methods, please refer to section 2. Evaluation is conducted using the most widely adopted metrics in this field: accuracy (Acc) and weighted F1 score (w-F1).

4.2 Performance Comparison

Table 1 presents the results on the IEMOCAP and MELD datasets. The suffix "gen" indicates that predicted historical emotions are used in the experiments; otherwise, groundtruth emotions are used. The comparison between them suggests that higher prediction accuracy of historical emotions leads to improved future predictions. And our proposed HRG-SSA outperformed all the baseline models. Compared to HAUCL, HRG-SSA achieves improvements of 6.29% in ACC and 6.48% in w-F1 on IEMOCAP, and 0.34% in ACC and 0.86% in w-F1 on MELD. The model shows greater improvement on IEMOCAP, likely due to its simpler setup with fewer emotion categories and two-person dialogues. In contrast, MELD is more complex, with challenges like imbalanced emotion labels, leading to smaller gains.

4.3 Ablation Study

To more comprehensively evaluate the effectiveness of our proposed HRG-SSA method, we conducted ablation experiments to examine the contributions of PSA, MSA, and IC, as well as the impact of each modality on emotion recognition accuracy. The results are presented in Table 4.

Methods	IEMOCAP		MELD	
	Acc	w-F1	Acc	w-F1
(Ours)	75.48	75.47	68.05	66.83
w/o PSA	73.94	73.93	67.82	66.66
w/o MSA	74.61	74.63	67.73	66.88
w/o IC	73.81	73.68	67.85	66.72
T	74.43	74.49	67.32	65.57
T + A	74.74	74.63	67.85	66.48
T + V	73.62	73.41	67.24	65.70

Table 4: Performance of HRG-SSA for ablation study. "w/o" denotes "without", while "T, A, V" represent the text, acoustic, and visual modalities, respectively.

In the table, "w/o PSA" and "w/o MSA" indicate the removal of the two losses, $\mathcal{L}_{PSA}$ and $\mathcal{L}_{MSA}$, respectively, while "w/o IC" denotes the exclusion of all implicit connections. The results demonstrate that removing any of the afore-

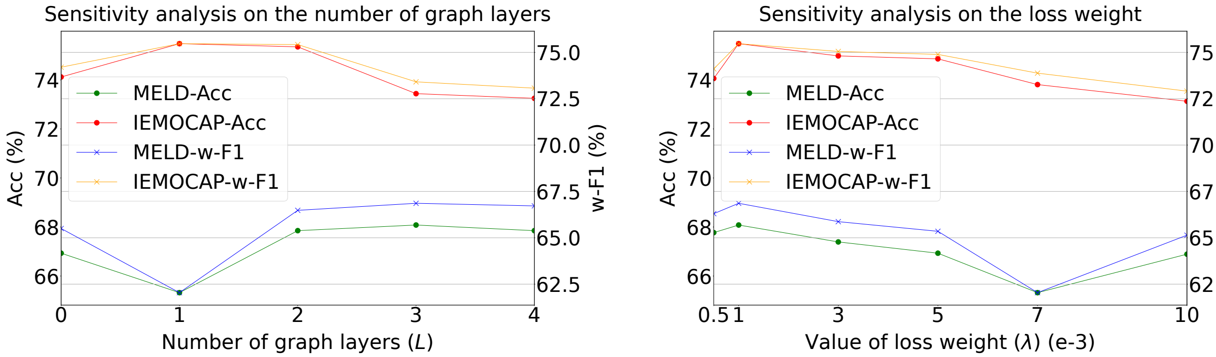


Figure 3: Sensitivity analysis of HRG-SSA on the λ and L. All experiments were conducted with other parameters fixed at optimal values.

mentioned modules leads to a decrease in emotion recognition accuracy, highlighting the positive contribution of each module. Moreover, removing PSA or MSA alone causes a significant performance drop, indicating that the two modules complement each other and validating the effectiveness of the progressive semantic alignment strategy.

The results in Table 4 indicate that the model performs well with only the text modality. Adding audio improves accuracy, while the inclusion of visual information reduces it. We believe that speech and text share some semantic overlap, enabling them to complement each other. However, the visual information in the datasets depicts a large scene, with the speaker's facial expression being only a small part of it. This may hinder the model's ability to focus effectively, leading to a decrease in accuracy. The collaboration of the three modalities results in a significant improvement, highlighting the subtle collaborative mechanism among them and offering us a more comprehensive and multidimensional perception.

4.4 Sensitivity Analysis

We selected two key parameters in HRG-SSA for sensitivity analysis on the IEMOCAP and MELD datasets, as illustrated in Figure 3. The left subfigure illustrate the impact of the number of stacked graph layers on model performance. Both graphs exhibit a trend of increasing and then decreasing performance, suggesting that both overly shallow and overly deep graph stacks hinder effective node information aggregation. Performance is notably poor when the number of layers is 0 (i.e., without the hybrid relationship graphs), emphasizing its importance. Additionally, the optimal performance of IEMOCAP and MELD is achieved at 1 and 3 layers, respectively. This can be attributed to the intrinsic properties of the datasets. IEMOCAP, which comprises two-person dialogues, inherently establishes implicit relationships through multi-hop node interactions, thus diminishing the necessity for deeper nesting. In contrast, MELD is a multi-party conversation dataset and contains a certain amount of invalid spoken content, raising the cost of information mining. The right subfigure show the effect of the loss weight λ and both datasets achieve the best performance at 0.001.

4.5 Visualization

Since we believe that text plays a key role in the MERC task, we take the example shown by Figure 1 in section 1, and draw a hybrid relationship graph constructed by its textual modality, as well as the distribution of attention among the nodes, as shown in Figure 4. The figure provides a clear visual representation of the model's mechanism. The red dashed line with arrows highlights the implicit connections mined by HRG-SSA, and "0" indicates no edges between nodes. Our model accurately identifies the full range of emotions by leveraging implicit connections and more rational attention allocation, while the baseline model HAUCL misclassifies U_5 as **Angry**.

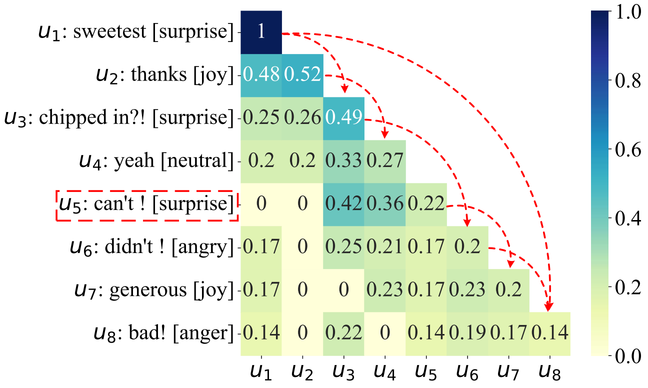


Figure 4: Heatmap of attention distribution among nodes in the hybrid relationship graph of the text modality constructed in the example shown in Figure 1. The horizontal and vertical axes are aligned, with text content omitted.

5 Conclusion

In this paper, we propose an end-to-end framework for MERC based on comparative learning and graphs. This framework aims to effectively utilize historical sentiment and emphasize implicit relationships within modalities. Our proposed method HRG-SSA complements the dialog graph by mining implicit relationships within each modality, thereby facilitating the alignment of sentiment-rich multimodal information. This improves the model's ability to extract sentiment cues and enhances accuracy. In the future, we will further explore the variability factors within each modality to further improve the performance of the model by constructing adaptive multimodal heterogeneous maps, as well as enhancing the interpretability of the model.

Acknowledgements

This research was supported by the National Natural Science Foundation of China (Nos. 62471403, 62261136549, 62271411, U22A2098), the Technological Innovation Team of Shaanxi Province (No. 2025RS-CXTD-009), the International Cooperation Project of Shaanxi Province (No. 2025GH-YBXM-017), the Fundamental Research Funds for the Central Universities (Nos. G2024WD0151, D5000240309).

References

[Ai *et al.*, 2024] Wei Ai, Yuntao Shou, Tao Meng, and Keqin Li. Der-gcn: Dialog and event relation-aware graph convolutional neural network for multimodal dialog emotion recognition. *IEEE Transactions on Neural Networks and Learning Systems*, 2024.

[Baltrušaitis *et al.*, 2019] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2):423–443, Feb 2019.

[Busso *et al.*, 2008] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42:335–359, 2008.

[Chen *et al.*, 2023] Feiyu Chen, Jie Shao, Shuyuan Zhu, and Heng Tao Shen. Multivariate, multi-frequency and multimodal: Rethinking graph neural networks for emotion recognition in conversation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10761–10770, 2023.

[Eyben *et al.*, 2010] Florian Eyben, Martin Wöllmer, and Björn Wolfgang Schüller. Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM international conference on Multimedia*, pages 1459–1462, 2010.

[Geetha *et al.*, 2024] A.V. Geetha, T. Mala, D. Priyanka, and E. Uma. Multimodal emotion recognition with deep learning: Advancements, challenges, and future directions. *Information Fusion*, 105:102218, may 2024.

[Ghosal *et al.*, 2019] Deepanway Ghosal, Navonil Majumder, Soujanya Poria, Niyati Chhaya, and Alexander Gelbukh. Dialoguegcn: A graph convolutional neural network for emotion recognition in conversation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019.

[Hu *et al.*, 2021a] Jingwen Hu, Yuchen Liu, Jinming Zhao, and Qin Jin. 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*, pages 5666–5675. Association for Computational Linguistics, 2021.

[Hu *et al.*, 2021b] Jingwen Hu, Yuchen Liu, Jinming Zhao, and Qin Jin. 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*, pages 5666–5675, 2021.

[Hu *et al.*, 2022a] Dou Hu, Xiaolong Hou, Lingwei Wei, Lianxin Jiang, and Yang Mo. 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, 2022.

[Hu *et al.*, 2022b] Guimin Hu, Ting-En Lin, Yi Zhao, Guangming Lu, Yuchuan Wu, and Yongbin Li. 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, 2022.

[Hu *et al.*, 2024] Ronglong Hu, Jizheng Yi, Lijiang Chen, and Ze Jin. Graph reconstruction attention fusion network for multimodal sentiment analysis. *IEEE Transactions on Industrial Informatics*, pages 1–10, 2024.

[Huang *et al.*, 2017] Gao Huang, Zhuang Liu, Der Maaten Laurens Van, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.

[Huang *et al.*, 2023] Changqin Huang, Junling Zhang, Xuemei Wu, Yi Wang, Ming Li, and Xiaodi Huang. Tefna: Text-centered fusion network with crossmodal attention for multimodal sentiment analysis. *Knowledge-Based Systems*, 269:110502, 2023.

[Hubert *et al.*, 2019] Tsai Yao-Hung Hubert, Shaojie Bai, Paul Pu Liang, Kolter J Zico, Morency Louis-Philippe, and Salakhutdinov Ruslan. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the conference. Association for computational linguistics. Meeting*, volume 2019, page 6558, 2019.

[Lee and Choi, 2021] Bongseok Lee and Yong Suk Choi. Graph based network with contextualized representations of turns in dialogue. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 443–455, 2021.

[Li *et al.*, 2023a] Bobo Li, Fei Hao, Lizi Liao, Yu Zhao, Chong Teng, Tat-Seng Chua, Donghong Ji, and Fei Li. Revisiting disentanglement and fusion on modality and context in conversational multimodal emotion recognition. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 5923–5934, 2023.

[Li *et al.*, 2023b] Zaijing Li, Ting-En Lin, Yuchuan Wu, Meng Liu, Fengxiao Tang, Ming Zhao, and Yongbin Li.

Unisa: Unified generative framework for sentiment analysis. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 6132–6142, 2023.

[Li *et al.*, 2024] Jiang Li, Xiaoping Wang, and Zhigang Zeng. Tracing intricate cues in dialogue: Joint graph structure and sentiment dynamics for multimodal emotion recognition. *arXiv preprint arXiv:2407.21536*, 2024.

[Lian *et al.*, 2023] Zheng Lian, Lan Chen, Licai Sun, Bin Liu, and Jianhua Tao. Gcnet: Graph completion network for incomplete multimodal learning in conversation. *IEEE Transactions on pattern analysis and machine intelligence*, 45(7):8419–8432, 2023.

[Ma *et al.*, 2023] Hui Ma, Jian Wang, Hongfei Lin, Bo Zhang, Yijia Zhang, and Bo Xu. A transformer-based model with self-distillation for multimodal emotion recognition in conversations. *IEEE Transactions on Multimedia*, 2023.

[Maji *et al.*, 2023] Bubai Maji, Monorama Swain, Rajlakshmi Guha, and Aurobinda Routray. Multimodal emotion recognition based on deep temporal features using cross-modal transformer and self-attention. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5, 2023.

[Nguyen *et al.*, 2023] Cam-Van Thi Nguyen, Anh-Tuan Mai, The-Son Le, Hai-Dang Kieu, and Duc-Trong Le. Conversation understanding using relational temporal graph neural networks with auxiliary cross-modality interaction. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15154–15167, 2023.

[Pei *et al.*, 2024] Hongbin Pei, Yu Li, Huiqi Deng, Jingxin Hai, Pinghui Wang, Jie Ma, Jing Tao, Yuheng Xiong, and Xiaohong Guan. Multi-track message passing: Tackling oversmoothing and oversquashing in graph learning via preventing heterophily mixing. In *Forty-first International Conference on Machine Learning*, 2024.

[Poria *et al.*, 2019] Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 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, 2019.

[Raffel *et al.*, 2020] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020.

[Rahman *et al.*, 2020] Wasifur Rahman, Md Kamrul Hasan, Sangwu Lee, Amir Zadeh, Chengfeng Mao, Louis-Philippe Morency, and Ehsan Hoque. Integrating multimodal information in large pretrained transformers. In *IEEE Transactions on Multimedia*, volume 2020, page 2359, 2020.

[Sun *et al.*, 2023] Licai Sun, Zheng Lian, Bin Liu, and Jianhua Tao. Efficient multimodal transformer with dual-level feature restoration for robust multimodal sentiment analysis. *IEEE Transactions on Affective Computing*, 15(1):309–325, 2023.

[Tu *et al.*, 2024a] Geng Tu, Jun Wang, Zhenyu Li, Shiwei Chen, Bin Liang, Xi Zeng, Min Yang, and Ruifeng Xu. Multiple knowledge-enhanced interactive graph network for multimodal conversational emotion recognition. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 3861–3874, 2024.

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

[Vaswani *et al.*, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, pages 5998–6008, 2017.

[Veličković *et al.*, 2018] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks. In *International Conference on Learning Representations*, 2018.

[Wang *et al.*, 2022] Di Wang, Shuai Liu, Quan Wang, Yumin Tian, Lihuo He, and Xinbo Gao. Cross-modal enhancement network for multimodal sentiment analysis. *IEEE Transactions on Multimedia*, 25:4909–4921, 2022.

[Yang *et al.*, 2019] Hao Yang, Chunfeng Yuan, Bing Li, Yang Du, Junliang Xing, Weiming Hu, and Stephen J Maybank. Asymmetric 3d convolutional neural networks for action recognition. *Pattern recognition*, 85:1–12, 2019.

[Yi *et al.*, 2024] Zijian Yi, Ziming Zhao, Zhishu Shen, and Tiehua Zhang. Multimodal fusion via hypergraph autoencoder and contrastive learning for emotion recognition in conversation. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 4341–4348, 2024.

[Zhang *et al.*, 2023a] Yazhou Zhang, Ao Jia, Bo Wang, Peng Zhang, Dongming Zhao, Pu Li, Yuexian Hou, Xiaojia Jin, Dawei Song, and Jing Qin. M3gat: A multi-modal, multi-task interactive graph attention network for conversational sentiment analysis and emotion recognition. *ACM Transactions on Information Systems*, 42(1), aug 2023.

[Zhang *et al.*, 2023b] Yazhou Zhang, Jinglin Wang, Yaochen Liu, Lu Rong, Qian Zheng, Dawei Song, Prayag Tiwari, and Jing Qin. A multitask learning model for multimodal sarcasm, sentiment and emotion recognition in conversations. *Information Fusion*, 93:282–301, 2023.

[Zhao *et al.*, 2022] Yu Zhao, Huaming Du, Ying Liu, Shaopeng Wei, Xingyan Chen, Fuzhen Zhuang, Qing Li, and Gang Kou. Stock movement prediction based on bi-typed hybrid-relational market knowledge graph via dual attention networks. *IEEE Transactions on Knowledge and Data Engineering*, 35(8):8559–8571, 2022.