Multiple Knowledge-Enhanced Interactive Graph Network for Multimodal Conversational Emotion Recognition

Geng Tu[1,2*], **Jun Wang**[1*], **Zhenyu Li**[1], **Shiwei Chen**[1,2], **Bin Liang**[3], **Xi Zeng**[6], **Min Yang**[4], **Ruifeng Xu**[1,2,5 †]

[1]Harbin Institute of Technology, Shenzhen, China [2]Peng Cheng Laboratory, China
[3]The Chinese University of Hong Kong, Hong Kong, China
[4]SIAT, Chinese Academy of Sciences, Shenzhen, China
[5]Guangdong Provincial Key Laboratory of Novel Security Intelligence Technologies
[6]The 30th Research Institute of China Electronics Technology Group Corporation, China
22b951011@stu.hit.edu.cn, 23s051031@stu.hit.edu.cn, xuruifeng@hit.edu.cn

Abstract

Multimodal Emotion Recognition in Conversations (ERC) aims to identify emotions in conversational videos. Current efforts focus on modeling both context-sensitive and speaker-sensitive dependencies and multimodal fusion. Despite the progress, models in Multimodal ERC (MERC) still struggle due to a lack of CommonSense Knowledge (CSK). In contrast, models in textual ERC typically employ CSK to enhance emotion inference. However, in multimodal scenarios, relying solely on textual CSK while neglecting visual CSK may hinder the understanding of visual emotional cues. To address this, we introduce a novel approach called Multiple Knowledge Enhanced Interactive Graph Network (MKE-IGN) to integrate multiple knowledge, such as textual and visual CSK, into the edge representations, thereby facilitating the modeling of relations between utterances and different types of CSK. Furthermore, considering that irrelevant CSK might be retained as noise, MKE-IGN adaptively selects this CSK guided by the mood-congruent effect and refines it based on contexts. Experimental results show that MKE-IGN outperforms state-of-the-art methods on two popular datasets.

1 Introduction

Emotion recognition in conversations (ERC) is challenging due to its dynamic and spontaneous nature, as individuals express various emotions (Zheng et al., 2023). Recently, ERC has garnered substantial interest because of its valuable applications in recommendation systems (Zou et al., 2022) and dialogue generation (Zhu et al., 2022; Saha et al., 2022; Lin et al., 2023).

Traditional ERC paradigms primarily focus on the textual modality and typically require modeling both context-sensitive and speaker-sensitive dependencies (Lian et al., 2021; Tu et al., 2023a; Wen et al., 2023b). Unfortunately, these methods struggle to replicate human-like understanding due to challenges in interpreting emotions expressed through Common Sense Knowledge (CSK) (Tu et al., 2023d). Recent research has addressed this limitation by modeling knowledge-sensitive dependencies into ERC models (Zhong and Di Wang, 2019; Tu et al., 2023b; Yang et al., 2023b,c) using knowledge bases such as ConceptNet (Speer et al., 2017), SenticNet (Cambria et al., 2022), and if-then CSK generated by the pre-trained commonsense transformer (COMET) (Bosselut et al., 2019). COMET is trained on ATOMIC (Sap et al., 2019), a collection of everyday inferential if-then commonsense knowledge organized through textual descriptions. This body of knowledge has significantly advanced the field of ERC.

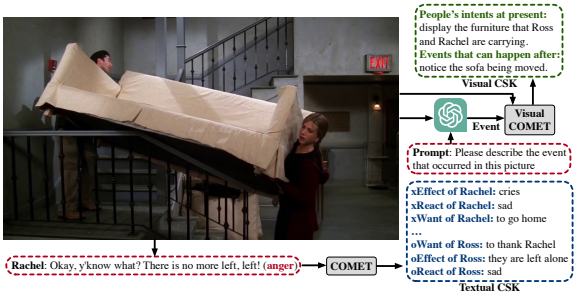

Figure 1: An example highlights the limitation of relying solely on textual if-then CSK for comprehending visual emotional cues in multimodal scenarios.

Despite the progress, textual cues prove insufficient for understanding deep emotions (Hazarika et al., 2018). Multimodal ERC (MERC), incorporating audio and visual cues alongside the text, is gaining increasing research attention (Yang et al., 2023a). Existing MERC methods focus on aggregation-based fusion through concatenation (Majumder et al., 2019; Ghosal et al., 2020b; Tu et al., 2022b), attention network (Shi and Huang, 2023; Yang et al., 2023a), and heterogeneous graph-based fusion (Yang et al., 2021; Hu et al., 2022;

[*] Equal contribution.
[†] Corresponding authors.

Findings of the Association for Computational Linguistics: EACL 2024, pages 3861–3874

Chen et al., 2023). However, these methods often rely on **future utterances to predict the current one's emotion**, which is impractical in real-world scenarios. More importantly, **integration of CSK in MERC has been overlooked**. Relying solely on textual if-then CSK can impede understanding of visual emotional cues, as it fails to incorporate visually contextualized information. For instance, visual if-then CSK can infer from an image that individuals moving a sofa will likely handle it with care later, a nuance that textual if-then CSK misses (see Fig. 1). Moreover, incorporating diverse types of if-then CSK inevitably introduces irrelevant CSK as noise (Tu et al., 2022a; Jiang et al., 2022; Tu et al., 2023c) in MERC. Yet, **research on the selection of multiple CSK still remains unexplored**.

To address the above issues, we propose an innovative solution, the Multiple Knowledge-Enhanced Interactive Graph Network (MKE-IGN), which employs a directed graph structure, eliminating the introduction of future utterances. It integrates multiple CSK (visual and textual if-then CSK) from VisualCOMET (Park et al., 2020) and COMET as the edge representations to enhance utterance understanding. This compensates for the deficiency in textual CSK regarding the comprehension of visual emotional clues. Especially, compared to incorporating knowledge into node representations, the edge representations in MKE-IGN enable the modeling of relations between utterances and various types of if-then CSK through diverse edge types. Furthermore, to reduce noise from irrelevant CSK, MKE-IGN leverages LassoNet (Lemhadri et al., 2021) to adaptively select the most emotionally relevant CSK from the generated candidate set, guided by the mood-congruent effect. Considering that if-then CSK is often regarded as a fixed, gradient-less representation, limiting its adaptability across various conversational contexts, we refine it based on contexts during training. Our main contributions can be summarized as follows:

- Pioneering the exploration of multiple if-then CSK in the MERC task.
- Introducing MKE-IGN, a novel solution that integrates multiple CSK into edge representations, surpassing state-of-the-art methods.
- Proposing an adaptive approach for the knowledge selection guided by the mood-congruent effect and the knowledge refinement based on contextual cues in multimodal scenarios.

2 Related Work

Emotion Recognition in Conversations

Context-sensitive Models: The emotion generation theory (Gross and Barrett, 2011) indicates the importance of contextual information for emotion identification. RNN-based models (Poria et al., 2017) are often used to model context dependencies. However, they are unable to capture the distinction between historical utterances (Lian et al., 2021) when modeling context. To solve this problem, most works began to focus on the memory network (Hazarika et al., 2018; Jiao et al., 2020). In addition, the role of participants in ERC is also important to the speaker's emotional state (Wen et al., 2023a). To model the **speaker-sensitive dependency**, researchers have a greater emphasis on speaker-specific models (Kim and Vossen, 2021), graph-based models (Nie et al., 2021), and so on. For example, Majumder et al. (2019) utilized three GRUs to track global context, speaker state, and emotional state in conversations. Guo et al. (2024) introduced a speaker-aware network to extract emotional cues by simulating cognitive conversational dynamics. Shen et al. (2021) and Lian et al. (2023) employed a graph-based model to model self- and inter-speaker dependencies.

Knowledge-sensitive Models: Although the above works have achieved respectable performance, they are not able to work like a human because of the lack of commonsense knowledge (Zhong and Di Wang, 2019). Recent research has addressed this limitation by modeling knowledge-sensitive dependencies into ERC models (Ghosal et al., 2020a; Fu et al., 2021; Li et al., 2021; Zhao et al., 2022; Tu et al., 2023b; Yang et al., 2023b,c) using knowledge bases such as ConceptNet (Speer et al., 2017), SenticNet (Cambria et al., 2022), and if-then CSK generated by pre-trained commonsense transformers (COMET) (Bosselut et al., 2019), which has significantly advanced the field of ERC. For instance, Fu et al. (2021) proposed a graph-based model to model the knowledge-sensitive dependencies by incorporating concepts retrieved from ConceptNet (Speer et al., 2017). Yang et al. (2023c) introduced if-then CSK into the conversation model through customized architectures, enhancing its emotional reasoning capabilities.

Multimodal Fusion: As multi-modality draws nearer to real-world application scenarios, MERC has been garnering growing research attention in recent years (Shi and Huang, 2023). Multimodal

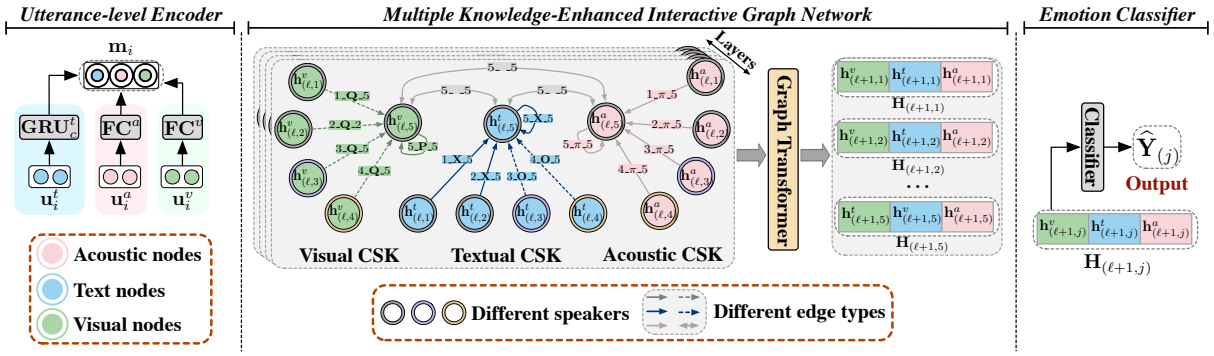

Figure 2: Illustration of MKE-IGN framework. The mathematical symbols align with the formulas in the paper text. The notation $i\_\mathbf{Q}\_j$ represents the edge between nodes i and j, with its edge representation denoted as $\mathbf{Q}$ of node i.

fusion in MERC aims to combine information from different modalities, including aggregation-based methods like concatenation (Hazarika et al., 2018; Tu et al., 2022b) and attention networks (Rahman et al., 2020; Wang et al., 2019). However, aggregation-based fusion methods overlook the complex interactions between modalities, resulting in insufficient utilization of contextual information (Hu et al., 2022). Recently, researchers have explored graph-based fusion methods to capture intra- and inter-modal interactive information (Hu et al., 2021b, 2022; Zhang and Li, 2023; Chen et al., 2023; Nguyen et al., 2023).

3 Methodology

In this section, we provide a detailed introduction to each component of the proposed MKE-IGN, as depicted in Fig. 2.

3.1 Task Definition

Let $\mathbf{U} = [\mathbf{u}_1, ..., \mathbf{u}_\mathbf{N}]$ be a conversation uttered by $\mathbf{M} \geq 2$ speakers, consisting of $\mathbf{N}$ utterances. Each utterance $\mathbf{u}_i$ is represented by a triplet $\mathbf{u}_i = \{\mathbf{u}_i^a, \mathbf{u}_i^v, \mathbf{u}_i^t\}$, where $\mathbf{u}_i^a \in \mathbb{R}^{d_a}$, $\mathbf{u}_i^v \in \mathbb{R}^{d_v}$, and $\mathbf{u}_i^t \in \mathbb{R}^{d_t}$ denote the acoustic, visual, and text features of $\mathbf{u}_i$, respectively. Multimodal ERC aims to predict the emotion label of each utterance $\mathbf{u}_i$.

3.2 Feature Representation

Utterance Feature Extraction[1]**:** Following Chen et al. (2023), we employ layer normalization and average operation on the last four hidden layers of the Roberta Large model (Liu et al., 2019) to obtain textual features. For extracting acoustic and visual features, we utilize OpenSmile (Schuller et al., 2011), an audio feature extraction toolkit, and a pre-trained DenseNet model (Huang et al., 2017) as per previous works (Hu et al., 2021b, 2022; Wen et al., 2023a; Jiang et al., 2023).

Knowledge Feature Extraction[2]**:** We employ the COMET model (Bosselut et al., 2019) to extract textual if-then CSK features, which is trained on the ATOMIC dataset (Sap et al., 2019) and categorizes if-then CSK into 9 relation types: *xWant*, *xIntent*, *xAttr*, *xNeed*, *xEffect*, *xReact*, *oWant*, *oEffect*, *oReact*. We utilize VisualCOMET (Park et al., 2020) trained on 1.4 million descriptions from 59,000 images. It extracts visual if-then CSK features categorized into 3 relation types: *Intent*, *before*, *after* using images and event descriptions obtained via Minigpt-4 (Zhu et al., 2023).

In textual if-then CSK, *xIntent* refers to the previous intent of speakers (wanted to do before) (Li et al., 2021), similar to the relation of *before* in visual if-then CSK. Thus, we avoid using them because they create dependencies between future utterances and the current utterance.

3.3 Utterance-level Encoder

Following (Hu et al., 2022), we employ a bidirectional GRU $\overleftrightarrow{\mathbf{GRU}}_c \in \mathbb{R}^{d_h \times d_t}$ to aggregate the contextual information for textual modality. The contextual representation $\widehat{\mathbf{m}}_i^t = \overleftrightarrow{\mathbf{GRU}}_c(\mathbf{u}_i^t, \widehat{\mathbf{h}}_{i-1}^t)$, where $\widehat{\mathbf{h}}_{i-1}^t$ is the hidden state. For acoustic and visual modality, we employ a fully connected layer $\mathbf{FC} \in \mathbb{R}^{d_h \times d_{a/v}}$ to obtain a consistent fixed-size representation $\mathbf{m}_i^\eta$, where $\eta \in \{a, v\}$.

3.4 Multiple Knowledge-Enhanced Interactive Graph Network

To integrate if-then CSK and its inherent relationships, we construct a directed graph where vari-

[1]Please refer to the Appendix A for more details.

[2]Please refer to Table. 5 in the Appendix for definitions of the relations in textual and visual if-then CSK.

ous types of these CSK serve as heterogeneous edges, representing different relations of connections. Here are the specific details.

3.4.1 Interactive Graph Network

Graph Structure: We suggest a multimodal directed graph $\mathbf{G} = \{\nu, \delta, \mathcal{R}, \mathcal{A}\}$ to prevent future utterances from influencing the emotional inference of the current utterance. $\mathbf{m}_i^\xi \in \nu$ is the graph node of the i-th uttreance under modality ξ. $\mathbf{r} \in \mathcal{R}$ is the edge type. $\alpha_{(i,j)}^\xi \in \mathcal{A}$ is the representation of the edge $\mathbf{e}_{(i,j)}^\xi = (\mathbf{m}_i^\xi, \mathbf{r}, \mathbf{m}_j^\xi) \in \delta$, where $\forall i \leq j$.

Relations: Textual if-then CSK $\alpha_{(i,j)}^t$ is divided into two categories: $\mathbf{X}$={*xWant*, *xAttr*, *xNeed*, *xEffect*, *xReact*} $\in \mathbb{R}^{5d_c}$ of speakers and $\mathbf{O}$={*oWant*, *oEffect*, *oReact*} $\in \mathbb{R}^{3d_c}$ of listeners. The relation of $\mathbf{u}_i^t \rightarrow \mathbf{u}_j^t$ is denoted as $\mathbf{X}$ of node i if they have the same speaker, otherwise as $\mathbf{O}$ of node i.

For the extracted **visual if-then CSK** $\alpha_{(i,j)}^v$, they are also divided into two categories: $\mathbf{P}$={*Intent*} $\in \mathbb{R}^{d_c}$ denotes the current speaker's intent, while $\mathbf{Q}$={*after*} $\in \mathbb{R}^{d_c}$ represents the subsequent intent of the speaker. The relation of $\mathbf{u}_i^v \rightarrow \mathbf{u}_j^v$ is denoted as $\mathbf{P}$ of node i if $i = j$, and as $\mathbf{Q}$ of node i if $i < j$.

Since the acoustic modality a lacks if-then CSK, we use a fully connected layer $\mathbf{FC_c}$ to combine nodes i and j generating $\pi \in \mathbb{R}^{d_c}$ for the alternative formulation of the **acoustic if-then CSK** $\alpha_{(i,j)}^a$. Therefore, the relation of $\mathbf{u}_i^a \rightarrow \mathbf{u}_j^a$ can be denoted as π of node i if $i \leq j$.

Additionally, the relation of cross-modal connections $\mathbf{u}_j^{\xi_1} \rightarrow \mathbf{u}_j^{\xi_2}$ is represented by a zero-filled vector $\mathbb{I} \in \mathbb{R}^{d_c}$, where $\xi_1, \xi_2 \in \{a, v, t\}$ and $\xi_1 \neq \xi_2$.

Node Representation Update: We employ an $\mathbf{N}_\ell$-layer graph transformer (Yun et al., 2019) to propagate interactive information, updating the node representation $\mathbf{h}_{(\ell,j)}^\xi \in \mathbf{R}^{d_h}$ at each layer ℓ.

$$\mathbf{h}_{(\ell+1,j)}^\xi = (1 - \mathbf{g}_j^\xi)\Big(\sum\nolimits_{i \in \mathcal{N}} \mathbf{\Gamma}_{(i,j)}^\Xi \mathbf{z}_i^\Xi\Big) + \mathbf{g}_j^\xi \mathbf{W}_h^\xi \mathbf{h}_{(\ell,j)}^\xi \quad (1)$$

where $\mathcal{N}$ denotes neighbor node indices of node j. $\mathbf{z}_i^\Xi$ is the passed message involving the selected $\alpha_{(\ell,i,j)}^\Xi$, denoted as $\mathbf{c}_{(\ell,i,j)}^\Xi \in \mathbb{R}^{d_h}$. Usually, $\xi = \Xi$, but not necessarily when $i = j$. $\mathbf{g}_j^\xi$ is the gate for the residual connections. $\mathbf{\Gamma}_{(i,j)}^\Xi$ is the attention score for gathering information.

$$\mathbf{z}_i^\Xi = \mathbf{FC}_z^\Xi\big(\mathbf{h}_{(\ell,i)}^\Xi\big) + \mathbf{W}_z^\Xi \mathbf{c}_{(\ell,i,j)}^\Xi \quad (2)$$

$$\mathbf{\Gamma}_{(i,j)}^\Xi = Softmax\Big(\mathbf{FC}_q^\xi\big(\mathbf{h}_{(\ell,j)}^\xi\big)\big(\mathbf{FC}_k^\Xi\big(\mathbf{h}_{(\ell,i)}^\Xi\big) + \mathbf{W}_h^\Xi \mathbf{c}_{(\ell,i,j)}^\Xi\big)/\sqrt{d_{head}}\Big) \quad (3)$$

where $\mathbf{W}_z^\Xi \in \mathbb{R}^{d_{head} \times d_h}$ and $\mathbf{W}_h^\Xi \in \mathbb{R}^{d_{head} \times d_h}$ are the trainable weights. $\mathbf{FC}_q^\xi$ and $\mathbf{FC}_k^\Xi$ are employed for projection to obtain sets of queries and keys. And $\mathbf{FC}_z^\Xi$ is a fully connected layer that maps from $\mathbb{R}^{d_h}$ to $\mathbb{R}^{d_{head}}$. Following (Li et al., 2021), we concatenate outputs from all heads to obtain $\mathbf{o}_j^\xi$. Therefore, $\mathbf{g}_j^\xi = sigmoid(\mathbf{W}_o^\mathbf{T}[\mathbf{h}_{(\ell,j)}^\xi; \mathbf{o}_j^\xi; \mathbf{h}_{(\ell,j)}^\xi - \mathbf{o}_j^\xi])$, where $[;]$ represents the concatenating operation. We denote the final output as $\mathbf{H}_{(\mathbf{N}_\ell,j)} = [\mathbf{h}_{(\mathbf{N}_\ell,j)}^a; \mathbf{h}_{(\mathbf{N}_\ell,j)}^v; \mathbf{h}_{(\mathbf{N}_\ell,j)}^t] \in \mathbb{R}^{3d_h}$.

3.4.2 Knowledge Selection

In GPT-based models (Radford et al., 2018) such as COMET and VisualCOMET, representations are typically generated sequentially. The hidden state of the final token is viewed as encapsulating the sequence's semantic information, serving as the representation for if-then CSK.

However, relying solely on the last token may not fully capture the semantic information from preceding tokens, potentially leading to incomplete understanding (Reimers and Gurevych, 2019). To address this, we integrate the last token with the previous 4 tokens, creating 5 candidates from which we dynamically select.

$$\mathbf{\Psi}_{(\kappa)}^\xi = tanh\big(\mathbf{W}_s^\xi \alpha_{(1,i,j,\kappa)}^\xi\big), \quad (4)$$

$$\mathbf{\Phi}_{(\kappa)}^\xi = Softmax\Big(\mathbf{\Psi}_{(\kappa)}^\xi \big(\mathbf{h}_{(\ell,i)}^\xi\big)^\mathbf{T}\Big) \quad (5)$$

$$\widehat{\mathbf{c}}_{(1,i,j)}^\xi = \mathbf{FC}_\kappa^\xi\Big(\sum\nolimits_{\kappa=1\ldots5} \mathbf{\Phi}_{(\kappa)}^\xi \alpha_{(1,i,j,\kappa)}^\xi\Big) \quad (6)$$

where $\alpha_{(1,i,j,\kappa)}^\xi \in \mathbb{R}^{d_c/3d_c/5d_c}$ is the κ-th candidate. $\mathbf{W}_s^\xi \in \mathbb{R}^{d_h \times d_c/3d_c/5d_c}$ is a trainable weight. $\mathbf{FC}_\kappa^\xi$ transforms the dimension from $\mathbb{R}^{d_c/3d_c/5d_c}$ to $\mathbb{R}^{d_h}$. Since acoustic CSK employs zero-filling, no knowledge selection is required.

Emotion-Aware Feature Selection: According to the mood-congruency effect (Mayer et al., 1990), individuals tend to select and process information that is congruent with their current emotional state, exhibiting a form of emotional priming. However, these if-then CSK are often too generic, making

it difficult for models to capture features highly relevant to emotion in knowledge representations.

Inspired by LassoNet (Lemhadri et al., 2021), which introduces Lasso regression-based characteristics (Tibshirani, 1996) into the framework of neural networks by adding an $\mathbf{L}_1$ regularization term to the loss function, resulting in some regression coefficients becoming zero, implying that the network can not only make predictions but also select features highly related to the target variable (such as emotional labels).

$$\mathbf{c}^{\xi}_{(1,i,j)}, \widetilde{\mathbf{Y}}^{\xi}_{i} = \mathbf{LassoNet}^{\xi}\big(\widehat{\mathbf{c}}^{\xi}_{(1,i,j)}\big) \quad (7)$$

where $\mathbf{c}^{\xi}_{(1,i,j)} \in \mathbb{R}^{d_h}$ is the edge representation after knowledge selection. $\widetilde{\mathbf{Y}}^{\xi} \in \mathbb{R}^{\mathbf{N}}$ is the predicting emotional label set for the utterance's CSK.

3.4.3 Knowledge Refinement

Previous studies (Tu et al., 2022a) often regarded if-then CSK as a fixed, gradient-less representation, limiting its adaptability across various conversational contexts. To better align knowledge with both the historical utterance i and the current utterance j, we have conducted further refinements.

$$\mathbf{c}^{\xi}_{(\ell+1,i,j)} = \mathbf{FC}^{\xi}_{\varrho}\big([\mathbf{p}^{\xi}_{(\ell,i,j)}; \Delta_i \mathbf{p}^{\xi}_{(\ell,i,j)}; \Delta_j \mathbf{p}^{\xi}_{(\ell,i,j)}]\big) \quad (8)$$

where $\mathbf{p}^{\xi}_{(\ell,i,j)} = \mathbf{FC}^{\xi}_{\rho}\big(\mathbf{c}^{\xi}_{(\ell,i,j)}\big)$. $\mathbf{FC}^{\xi}_{\varrho} \in \mathbb{R}^{d_h \times 3d_{head}}$ and $\mathbf{FC}^{\xi}_{\rho} \in \mathbb{R}^{d_{head} \times d_h}$ are used for projection dimension. Δ represents the attention score, computed similarly to Formula (3).

3.5 Emotion Classifier

We utilize a linear unit to predict the emotion distributions:

$$\widehat{\mathbf{Y}}_{(j)} = Argmax(Softmax(\mathbf{W}_e \mathbf{H}_{(\mathbf{N}_\ell,j)} + \mathbf{b}_e)) \quad (9)$$

where $\mathbf{W}_e \in \mathbb{R}^{d_e \times 3d_h}$ and $\mathbf{b}_e \in \mathbb{R}^{d_e}$ are trainable parameters, d_e is the number of emotion categories. $\widehat{\mathbf{Y}} \in \mathbb{R}^{\mathbf{N}}$ is the predicting emotional label set of utterances in a conversation.

$$\mathcal{L}_{all} = \mathcal{L}^{v} + \mathcal{L}^{t} + \mathcal{L}_{ce} \quad (10)$$

$$\mathcal{L}_{ce} = CrossEntropy(\widehat{\mathbf{Y}}, \mathbf{Y}) + \varsigma \left\|\Theta\right\|_2 \quad (11)$$

$$\mathcal{L}^{\xi} = CrossEntropy(\widetilde{\mathbf{Y}}^{\xi}, \mathbf{Y}) + \lambda||\Theta^{\xi}||_1 \quad (12)$$

where $\mathcal{L}_{ce}$ is the classification loss. $\mathbf{Y} \in \mathbb{R}^{\mathbf{N}}$ represents the set of true labels. Θ and Θ^{ξ} are the set of projection parameters. ς and λ denotes the coefficient of $\mathbf{L}_2$ and $\mathbf{L}_1$ regularization, respectively.

Dataset	Dialogues			Utterances			Classes
	train	val	test	train	val	test	
MELD	1039	114	280	9,989	1,109	2610	7
IEMOCAP	120		31	5,810		1,623	6

Table 1: Statistics of two conversational datasets.

4 Experiments

4.1 Experimental Data and Settings

Datasets: We benchmark MKE-IGN on two datasets: **IEMOCAP** (Busso et al., 2008) has dyadic conversation videos with ten speakers, featuring 7,433 utterances and 151 dialogues. Each utterance has one of six emotions. **MELD** (Poria et al., 2018) contains multiparty conversations collected from the 'Friends' TV series, having 1,433 conversations, 13,708 utterances, and 304 speakers. Each utterance holds one of seven emotions. Following (Ghosal et al., 2020a), the data splitting for datasets is detailed in Table 1. As the IEMOCAP lacks a predefined train/validation split, we allocate 10% of the training dialogues for validation.

Settings: We conduct a hyperparameter search for MKE-IGN using the validation set on each dataset. For IEMOCAP, the model uses a learning rate of 5e-5 with an AdamW optimizer and a batch size of 8. The graph transformer has node dimension $d_h = 200$, head dimension $d_{head} = 50$, and 2 layers ($\mathbf{N}_\ell = 2$). For MELD, the learning rate is 1e-4, batch size is 16, $d_h = 512$, $d_{head} = 256$, and 5 layers ($\mathbf{N}_\ell = 5$). All experiments are conducted on a single GeForce RTX 4090 GPU and reported results are averages from 5 random test set runs.

4.2 Comparison Methods

To comprehensively evaluate MKE-IGN, we compare it with the following state-of-the-art methods: **DialogueRNN** (Majumder et al., 2019) uses three GRUs for tracking speakers and context, while **DialogueGCN** (Ghosal et al., 2020b) tackles context propagation through a graph network; both use concatenated multimodal features. **CTNet** (Zhang et al., 2020) utilizes a transformer-based structure to model intra- and inter-modal interaction. **SCMM** (Yang et al., 2023a) combines context modeling, modal interaction, and path selection for enhanced multimodal features. **SACCMA** (Guo et al., 2024) integrates speaker-aware cognitive processing with cross-modal attention fusion to capture emotional cues. **MMDFN** (Hu et al., 2022) utilizes a heterogeneous graph to represent relation-

Methods	IEMOCAP								MELD								
	Happy	Sad	Neutral	Angry	Excited	Frustrated	Acc	w-F1	Neutral	Surprise	Fear	Sadness	Joy	Disgust	Anger	Acc	w-F1
DialogueRNN♢	32.20	80.26	57.89	62.82	73.87	59.76	63.52	62.89	76.97	47.69	-	20.41	50.92	-	45.52	60.31	57.66
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
CTNet♠	51.30	79.90	65.80	67.20	**78.70**	58.80	67.60	67.50	77.40	52.70	10.00	32.50	56.00	11.20	44.60	-	60.50
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
SCMM♠	45.37	78.76	63.54	66.05	76.70	66.18	-	67.53	-	-	-	-	-	-	-	-	59.44
CMCF-SRNet♠	52.20	80.90	68.80	70.30	76.70	61.60	-	69.60	-	-	-	-	-	-	-	-	62.30
M^{3}Net∋	52.74	79.39	67.55	69.30	74.39	66.58	-	69.24	79.31	58.76	**20.51**	**40.46**	63.21	26.17	52.53	-	65.47
CORECT♠	**59.30**	80.53	66.94	69.59	72.69	68.50	69.93	70.02	-	-	-	-	-	-	-	-	-
SACCMA♠	38.60	**86.53**	64.90	64.56	74.52	62.99	-	67.10	-	-	-	-	-	-	-	-	59.30
IGN (Baseline)	51.03	79.58	66.92	63.08	75.29	63.06	67.78	67.66	78.53	57.73	14.93	35.83	62.43	24.24	51.95	65.40	64.26
MKE-IGN (Ours)	53.87	82.86	**72.07**	**71.30**	75.78	**68.84**	**72.03**	**71.93**	**80.00**	**59.82**	17.39	40.12	**64.01**	**31.30**	**56.08**	**67.78**	**66.56**

Table 2: Comparison of results (%) under the multimodal setting (acoustic, visual, and textual modalities). ♢, ∋, ♠ results come from (Hu et al., 2022), (Shi et al., 2023), and original papers, respectively. IGN refers to a variant of MKE-IGN that does not utilize any if-then CSK, replacing it with a zero-filled vector.

Methods	IEMOCAP		MELD	
	Acc	w-F1	Acc	w-F1
MKE-IGN (Ours)	72.03	71.93	67.78	66.56
w/o Visual CSK	70.86	70.64	66.13	64.91
w/o Textual CSK	69.62	69.81	66.70	64.69
w/o Acoustic CSK	70.98	70.98	66.28	65.34
w/o Knowledge Selection	70.30	70.09	66.28	64.78
w/o Knowledge Refinement	70.43	70.23	65.17	64.94

Table 3: Ablation results of MKE-IGN.

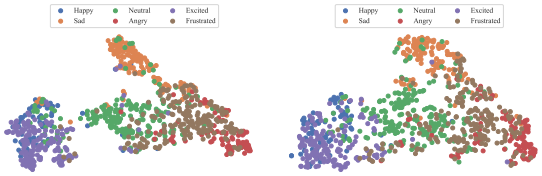

Figure 3: Visualization of intermediate utterance representations of MKE-IGN (left) and MKE-IGN w/o all CSK (right) on the IEMOCAP dataset.

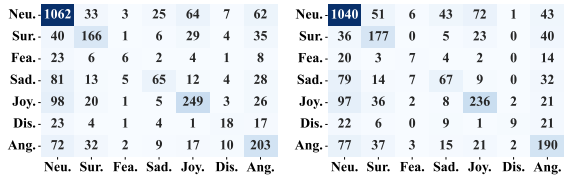

Figure 4: Confusion matrices of MKE-IGN (left) and MKE-IGN w/o visual CSK (right) on the MELD dataset.

ships between modalities. **MMGCN** (Hu et al., 2021b) employs a graph-based fusion module for capturing both intra- and inter-modal contextual features. **CMCF-SRNet** (Zhang and Li, 2023) integrates cross-modal interaction through a locality-constrained transformer and improves semantic relationships between utterances using a graph-based semantic refinement transformer. **M^{3}Net** (Chen et al., 2023) utilizes a GNN to capture emotion relationships and assesses their importance with multi-frequency signals. **CORECT** (Nguyen et al., 2023) is a relational temporal graph neural network designed to capture cross-modality interactions and temporal dependencies.

4.3 Overall Results

Following (Nguyen et al., 2023), we utilize the Accuracy (Acc) and weighted F1 score (w-F1) as evaluation metrics. Table 2 reports a comparison of MKE-IGN against other models, including advanced graph-based ones such as CORECT, CMCF-SRNet, and M^{3}Net. The results demonstrate MKE-IGN's superior performance, establishing a new state-of-the-art benchmark. This further underscores the effectiveness of integrating multiple if-then CSK, along with knowledge selection and refinement policies, for the MERC task.

4.4 Ablation Study

To investigate the impact of each component within MKE-IGN, we conduct an ablation study, revealing significant performance improvements as shown in Table 3. Further statistical analysis supports this, with a p-value $\ll 0.05$ for the paired t-test.

Analysis of Multiple CSK: Table 3 demonstrates that the removal of any type of CSK results in reduced overall model performance. Fig. 3 illustrates how integrating external if-then CSK enhances the distinction between utterance representations from different categories. This supplementary CSK fills contextual gaps in the dataset that the model might miss, facilitating a deeper understanding of specific contexts and implicit information within them.

After removing textual CSK, Acc drops by 1.08% and 2.41%, while w-F1 decreases by 1.87% and 2.12% on MEDL and IEMOCAP datasets, re-

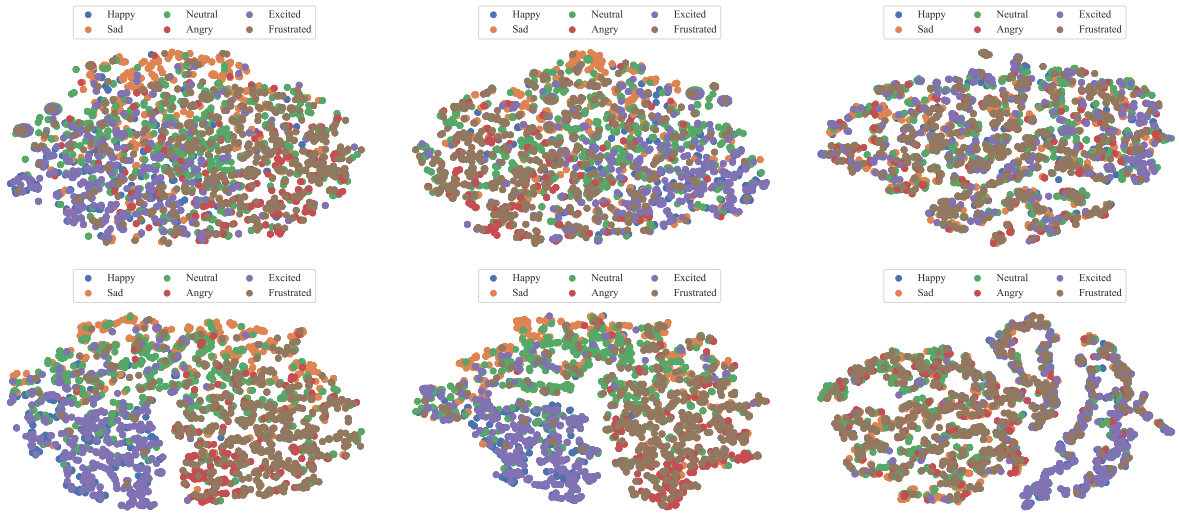

Figure 5: Visualization of intermediate embeddings of textual if-then CSK (**X**, **O**) and visual if-then CSK (first row), and these CSKs after knowledge selection (second row) on the IEMOCAP dataset.

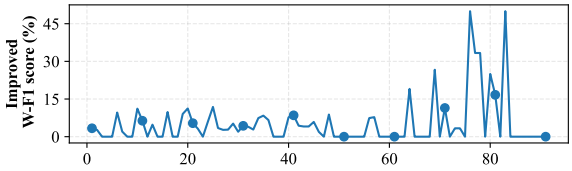

Figure 6: Lifting performance of MKE-IGN w/ Knowledge Refinement on different positions in conversations on the IEMOCAP dataset.

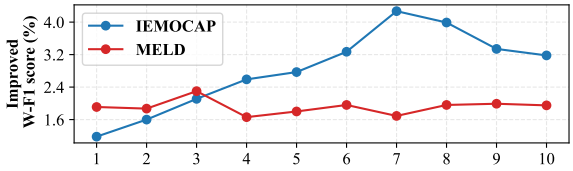

Figure 7: Lifting performance of MKE-IGN compared to the Baseline across different context window sizes.

spectively. Visual CSK performed worse than textual CSK on the IEMOCAP dataset due to the lack of visual scene information in the controlled recording studio setup. The significant performance boost on the MELD dataset, including rich visual scenes, further confirms this observation. In Fig. 4, Visual CSK demonstrates effectiveness across nearly all category samples, whereas 'Surprise', 'Fear', and 'Sadness' show a slight performance decrease, possibly influenced by the impact of class imbalance, leading to misclassification into 'Neutral' categories. Removing acoustic CSK could slightly decrease performance as it disrupts the graph's balance, potentially biasing the model's handling of graph data.

Analysis of Knowledge Selection: Table 3 highlights the improved performance achieved through knowledge selection. Specifically, Acc and w-F1 exhibit improvements of 1.73% and 1.84%, respectively, on the IEMOCAP dataset, and gains of 1.5% and 1.78%, respectively, on the MELD dataset. To explore how CSK changes after selection, we visualized the representations of textual and visual if-then CSK from the IEMOCAP dataset, as shown in Fig. 5. We observed that the if-then CSK processed through the knowledge selection effectively captures semantic distinctions among different emotion categories. This mitigates the noise introduced by these if-then CSK and enhances the model's emotional understanding in the MERC task.

Analysis of Knowledge Refinement: In addition to knowledge selection, knowledge refinement has also led to significant performance improvements, on IEMOCAP, Acc, and w-F1 improved by 1.6% and 1.7% respectively. On MELD, gains are 2.61% and 1.62%. Fig. 6 shows that performance improvements become significant after the 60th utterance. These enhancements mainly occur toward the end of conversations, likely because MERC models prioritize nearby context due to similar semantics (Ghosal et al., 2021). This makes it challenging to detect utterances requiring longer contexts. By enhancing its ability to adjust to various contexts while minimizing the influence of irrelevant CSK as background noise, the model can consider semantic elements from greater distances, which is crucial for handling long conversation data.

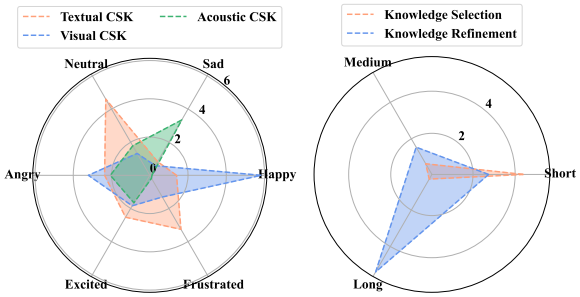

Figure 8: Lifting performance of each component of MKE-IGN across different emotional categories and conversation positions (Short: first 1/3, Medium: middle 1/3, and Long: last 1/3) on the IEMOCAP dataset.

4.5 Analysis of Variants of MKE-IGN

In Fig. 7, we demonstrate improved performance on the different datasets through the adjustment of context window sizes $\mathbf{N}_\omega$. Increasing $\mathbf{N}_\omega$ improves the model performance in lengthy dialogues like IEMOCAP, up to a point. Conversely, shorter conversations like MELD peak with a smaller window ($\mathbf{N}_\omega$=3), followed by fluctuations.

4.6 Complementarity Analysis

In this section, we explore the complementarity among three CSKs or knowledge selection and refinement to elucidate the rationality of MKE-IGN. **Complementarity of Multiple CSK:** Fig. 8 shows textual CSK performs well overall but less effectively in 'happy' and 'angry' minority categories compared to visual CSK, which contrasts with findings from the MELD dataset in Fig. 4. Surprisingly, acoustic CSK notably improves in 'sad', contrasting with the weaker performances of the other CSK in this category. **Complementarity of Knowledge Selection and Refinement:** Furthermore, in Fig. 8, we observe that knowledge selection primarily impacts the early stages of conversations. Performance declines in later stages, possibly due to fixed cognitive skills that struggle to adapt to more complex contexts. Knowledge refinement enhances adaptability to contexts, minimizing irrelevant background noise and enabling consideration of distant semantic elements crucial for handling long conversations, as illustrated in Fig. 6.

4.7 Comparing Results Against Large Language Models (LLMs)

Table 4 compares our method with various LLMs in the MERC task, outperforming zero-shot methods like ChatGPT[3], ChatGLM2 (Du et al., 2022), and Llama2 (Touvron et al., 2023). Even with a 3-shot performance, ChatGPT remains behind our method. Unlike these text-only approaches, we developed a prompt template for GPT-4o (as shown in Fig. 9) that supports text and video inputs, achieving scores of 44.15 on IEMOCAP and 62.21 on MELD. Fine-tuned models like LORA (Hu et al., 2021a) + Backbone also fall short. The lower performance of LLMs may be due to a mismatch between their interpretations and the specific labeling protocols of datasets (Zhao et al., 2023). This misalignment could explain why LLMs perform worse on scripted datasets like IEMOCAP compared to more naturalistic ones like MELD.

Methods	IEMOCAP	MELD
Ours	71.93	66.56
Zero-shot		
ChatGPT ♣	40.07	54.37
GPT-4o	44.15	62.21
ChatGLM ♦	38.60	38.80
ChatGLM2 ♦	21.10	21.80
Llama ♦	0.75	9.12
Llama2 ♦	2.77	16.28
Few-shot		
ChatGPT 1-shot ■	47.46	58.63
ChatGPT 3-shot ■	48.58	58.35
LORA + Backbone		
ChatGLM ♦	18.94	40.54
ChatGLM2 ♦	52.88	64.85
Llama ♦	55.81	66.15
Llama2 ♦	55.96	65.84

Table 4: Comparison of results against various LLMs. ♣, ♦ and ■ results are from (Tu et al., 2023d), (Lei et al., 2023) and (Zhao et al., 2023), respectively.

4.8 Error Analysis

Many errors in our method stem from class imbalance, as evidenced by the low F1 scores of 17.39% and 31.30% for the 'Fear' and 'Disgust' emotions, respectively, in the MELD dataset. This phenomenon also constitutes a primary constraint on the performance of the MERC task, a fact supported by the results in Table 2. Furthermore, we are also investigating cases where MKE-IGN misclassifies samples that IGN predicted correctly, totaling 93 samples in the IEMOCAP dataset. It notably struggles more with Short (50 samples) and Medium (34 samples) conversation positions, showing relatively better performance with Long positions (only 9 samples). This is likely due to the benefits of knowledge refinement.

[3]https://chat.openai.com/

Below are *N* utterances of a conversation. Each utterance contains text content and three temporally uniformly sampled frames.

utt_idx: utt_1
<UTT_1_SPEAKER>: <UTT_1_TEXT>
<UTT_1_IMG_1>
<UTT_1_IMG_2>
<UTT_1_IMG_3>

utt_idx: utt_2
<UTT_2_SPEAKER>: <UTT_2_TEXT>
<UTT_2_IMG_1>
<UTT_2_IMG_2>
<UTT_2_IMG_3>

…

utt_idx: utt_*N*
<UTT_*N*_SPEAKER>: <UTT_*N*_TEXT>
<UTT_*N*_IMG_1>
<UTT_*N*_IMG_2>
<UTT_*N*_IMG_3>

Please predict the emotion of each utterance based on the text and video content of historical conversations. Emotion must be selected from happy, sad, neutral, angry, excited, and frustrated. Return your answer in the form of a JSON dict. The key of the dict is utt_idx and the value is the predicted emotion.

Figure 9: The GPT-4o prompt template is tailored for the IEMOCAP dataset, with the flexibility to adapt to the MELD dataset by modifying the emotion categories.

5 Conclusion

This paper introduces a novel MKE-IGN for the MERC task, employing a directed graph structure to prevent future utterances from influencing the current one. More importantly, MKE-IGN integrates multiple CSK as edge representations for the modeling of relations between utterances and different types of CSK, improving the understanding of utterances in multimodal scenarios. Furthermore, considering that irrelevant CSK might be retained as noise, MKE-IGN adaptively selects this CSK guided by the mood-congruent effect and refines it based on contexts. Extensive evaluations and an ablation study prove the superiority of MKE-IGN and the significant impact of its components.

Limitations

While our proposed approach, MKE-IGN, demonstrates significant advancements in MERC, some limitations remain worth acknowledging. Firstly, despite incorporating multiple forms of CSK, including visual and textual if-then CSK, our model may not fully capture the intricacies of human emotional understanding. Emotions are complex and nuanced, often influenced by various contextual factors beyond the scope of available CSK. Secondly, the dynamic selection and refinement of CSK based on contextual cues aim to reduce noise from irrelevant knowledge. However, the effectiveness of this approach may vary depending on the quality and relevance of the contextual information. In certain cases, contextual cues may not adequately guide the selection process, leading to suboptimal performance. Lastly, while MKE-IGN outperforms existing methods on popular datasets, its generalizability to diverse conversational contexts and cultural nuances remains to be thoroughly explored. Emotions and their expressions can vary significantly across different cultures and social contexts, posing challenges for emotion recognition models trained on limited datasets. In summary, while MKE-IGN represents a promising step forward in multimodal emotion recognition, further research is needed to address the aforementioned limitations and enhance the model's robustness and applicability across various real-world scenarios.

Acknowledgements

We thank all anonymous reviewers for their helpful comments. This work was supported in part by the National Natural Science Foundation of China under Grants 62176076, the Natural Science Foundation of Guangdong under Grant 2023A1515012922, Shenzhen Foundational Research Funding under Grant JCYJ20220818102415032, The Major Key Project of PCL under Project PCL2023A09, and Guangdong Provincial Key Laboratory of Novel Security Intelligence Technologies under Grant 2022B1212010005k.

References

Emad Barsoum, Cha Zhang, Cristian Canton Ferrer, and Zhengyou Zhang. 2016. Training deep networks for facial expression recognition with crowd-sourced label distribution. In *Proceedings of the 18th ACM international conference on multimodal interaction*, pages 279–283.

Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi. 2019. Comet: Commonsense transformers for knowledge graph construction. In *ACL*.

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe

Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. 2008. Iemocap: Interactive emotional dyadic motion capture database. *LREC*, 42(4):335–359.

Erik Cambria, Qian Liu, Sergio Decherchi, Frank Xing, and Kenneth Kwok. 2022. Senticnet 7: A commonsense-based neurosymbolic ai framework for explainable sentiment analysis. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3829–3839.

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

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2022. Glm: General language model pretraining with autoregressive blank infilling. In *Proceedings of ACL*, pages 320–335.

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

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

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

Deepanway Ghosal, Navonil Majumder, Rada Mihalcea, and Soujanya Poria. 2021. Exploring the role of context in utterance-level emotion, act and intent classification in conversations: An empirical study. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1435–1449.

Deepanway Ghosal, Navonil Majumder, Soujanya Poria, Niyati Chhaya, and Alexander Gelbukh. 2020b. Dialoguegcn: A graph convolutional neural network for emotion recognition in conversation. In *EMNLP-IJCNLP*.

James J Gross and Lisa Feldman Barrett. 2011. Emotion generation and emotion regulation: One or two depends on your point of view. *Emotion Review*, 3(1).

Lili Guo, Yikang Song, and Shifei Ding. 2024. Speaker-aware cognitive network with cross-modal attention for multimodal emotion recognition in conversation. *Knowledge-Based Systems*, page 111969.

Devamanyu Hazarika, Soujanya Poria, Rada Mihalcea, Erik Cambria, and Roger Zimmermann. 2018. Icon: Interactive conversational memory network for multimodal emotion detection. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pages 2594–2604.

Dou Hu, Xiaolong Hou, Lingwei Wei, Lianxin Jiang, and Yang Mo. 2022. Mm-dfn: Multimodal dynamic fusion network for emotion recognition in conversations. In *ICASSP*, pages 7037–7041. IEEE.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021a. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Jingwen Hu, Yuchen Liu, Jinming Zhao, and Qin Jin. 2021b. Mmgcn: Multimodal fusion via deep graph convolution network for emotion recognition in conversation. In *ACL*, pages 5666–5675.

Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. 2017. Densely connected convolutional networks. In *CVPR*, pages 4700–4708.

Dazhi Jiang, Hao Liu, Runguo Wei, and Geng Tu. 2023. Csat-ftcn: A fuzzy-oriented model with contextual self-attention network for multimodal emotion recognition. *Cognitive Computation*, pages 1–10.

Dazhi Jiang, Runguo Wei, Jintao Wen, Geng Tu, and Erik Cambria. 2022. Automl-emo: Automatic knowledge selection using congruent effect for emotion identification in conversations. *IEEE Transactions on Affective Computing*.

Wenxiang Jiao, Michael Lyu, and Irwin King. 2020. Real-time emotion recognition via attention gated hierarchical memory network. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 8002–8009.

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

Shanglin Lei, Guanting Dong, Xiaoping Wang, Keheng Wang, and Sirui Wang. 2023. Instructerc: Reforming emotion recognition in conversation with a retrieval multi-task llms framework. *arXiv preprint arXiv:2309.11911*.

Ismael Lemhadri, Feng Ruan, Louis Abraham, and Robert Tibshirani. 2021. Lassonet: A neural network with feature sparsity. *Journal of Machine Learning Research*, 22(127):1–29.

Jiangnan Li, Zheng Lin, Peng Fu, and Weiping Wang. 2021. Past, present, and future: Conversational emotion recognition through structural modeling of psychological knowledge. In *Findings of the association*

for computational linguistics: EMNLP 2021, pages 1204–1214.

Zheng Lian, Lan Chen, Licai Sun, Bin Liu, and Jianhua Tao. 2023. Gcnet: Graph completion network for incomplete multimodal learning in conversation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Zheng Lian, Bin Liu, and Jianhua Tao. 2021. Ctnet: Conversational transformer network for emotion recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:985–1000.

Hsien-Chin Lin, Shutong Feng, Christian Geishauser, Nurul Lubis, Carel van Niekerk, Michael Heck, Benjamin Ruppik, Renato Vukovic, and Milica Gasić. 2023. Emous: Simulating user emotions in task-oriented dialogues. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2526–2531.

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

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

John D Mayer, Michael Gayle, Mary Ellen Meehan, and Anna-Kristina Haarman. 1990. Toward better specification of the mood-congruency effect in recall. *Journal of Experimental Social Psychology*, 26(6):465–480.

Cam Van Thi Nguyen, Tuan Mai, Dang Kieu, Duc Trong Le, et al. 2023. 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.

Weizhi Nie, Rihao Chang, Minjie Ren, Yuting Su, and Anan Liu. 2021. I-gcn: Incremental graph convolution network for conversation emotion detection. *IEEE Transactions on Multimedia*, pages 4471–4481.

Jae Sung Park, Chandra Bhagavatula, Roozbeh Mottaghi, Ali Farhadi, and Yejin Choi. 2020. Visualcomet: Reasoning about the dynamic context of a still image. In *ECCV*, pages 508–524. Springer.

Soujanya Poria, Erik Cambria, Devamanyu Hazarika, Navonil Majumder, Amir Zadeh, and Louis-Philippe Morency. 2017. Context-dependent sentiment analysis in user-generated videos. In *Proceedings of the 55th annual meeting of the association for computational linguistics (volume 1: Long papers)*, pages 873–883.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2018. Meld: A multimodal multi-party dataset for emotion recognition in conversations. *arXiv preprint arXiv:1810.02508*.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training.

Wasifur Rahman, Md Kamrul Hasan, Sangwu Lee, Amir Zadeh, Chengfeng Mao, Louis-Philippe Morency, and Ehsan Hoque. 2020. Integrating multimodal information in large pretrained transformers. In *ACL*, volume 2020, page 2359. NIH Public Access.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992.

Tulika Saha, Vaibhav Gakhreja, Anindya Sundar Das, Souhitya Chakraborty, and Sriparna Saha. 2022. Towards motivational and empathetic response generation in online mental health support. In *Proceedings of the 45th international ACM SIGIR conference on research and development in information retrieval*, pages 2650–2656.

Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A Smith, and Yejin Choi. 2019. Atomic: An atlas of machine commonsense for if-then reasoning. In *AAAI*, volume 33, pages 3027–3035.

Björn Schuller, Anton Batliner, Stefan Steidl, and Dino Seppi. 2011. Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge. *Speech Communication*, 53(9-10):1062–1087.

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

Tao Shi and Shao-Lun Huang. 2023. Multiemo: An attention-based correlation-aware multimodal fusion framework for emotion recognition in conversations. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14752–14766.

Tao Shi, Xiao Liang, Yaoyuan Liang, Xinyi Tong, and Shao-Lun Huang. 2023. Sslcl: An efficient model-agnostic supervised contrastive learning framework for emotion recognition in conversations. *arXiv preprint arXiv:2310.16676*.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31.

Robert Tibshirani. 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 58(1):267–288.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Geng Tu, Ran Jing, Bin Liang, Min Yang, Kam-Fai Wong, and Ruifeng Xu. 2023a. A training-free debiasing framework with counterfactual reasoning for conversational emotion detection. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15639–15650.

Geng Tu, Bin Liang, Dazhi Jiang, and Ruifeng Xu. 2022a. Sentiment-emotion-and context-guided knowledge selection framework for emotion recognition in conversations. *IEEE Transactions on Affective Computing*.

Geng Tu, Bin Liang, Xiucheng Lyu, Lin Gui, and Ruifeng Xu. 2023b. Do topic and causal consistency affect emotion cognition? a graph interactive network for conversational emotion detection. In *In The 26th European Conference on Artificial Intelligence (ECAI'23)*, pages 2362–2369.

Geng Tu, Bin Liang, Ruibin Mao, Min Yang, and Ruifeng Xu. 2023c. Context or knowledge is not always necessary: A contrastive learning framework for emotion recognition in conversations. In *Findings of ACL*, pages 14054–14067.

Geng Tu, Bin Liang, Bing Qin, Kam-Fai Wong, and Ruifeng Xu. 2023d. An empirical study on multiple knowledge from chatgpt for emotion recognition in conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12160–12173.

Geng Tu, Jintao Wen, Hao Liu, Sentao Chen, Lin Zheng, and Dazhi Jiang. 2022b. Exploration meets exploitation: Multitask learning for emotion recognition based on discrete and dimensional models. *Knowledge-Based Systems*, 235:107598.

Yansen Wang, Ying Shen, Zhun Liu, Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. 2019. Words can shift: Dynamically adjusting word representations using nonverbal behaviors. In *AAAI*, volume 33, pages 7216–7223.

Jintao Wen, Dazhi Jiang, Geng Tu, Cheng Liu, and Erik Cambria. 2023a. Dynamic interactive multiview memory network for emotion recognition in conversation. *Information Fusion*, 91:123–133.

Jintao Wen, Geng Tu, Rui Li, Dazhi Jiang, and Wenhua Zhu. 2023b. Learning more from mixed emotions: A label refinement method for emotion recognition in conversations. *Transactions of the Association for Computational Linguistics*, 11:1485–1499.

Haozhe Yang, Xianqiang Gao, Jianlong Wu, Tian Gan, Ning Ding, Feijun Jiang, and Liqiang Nie. 2023a. Self-adaptive context and modal-interaction modeling for multimodal emotion recognition. In *Findings of ACL*, pages 6267–6281.

Jianing Yang, Yongxin Wang, Ruitao Yi, Yuying Zhu, Azaan Rehman, Amir Zadeh, Soujanya Poria, and Louis-Philippe Morency. 2021. Mtag: Modal-temporal attention graph for unaligned human multimodal language sequences. In *NAACL*, pages 1009–1021.

Kailai Yang, Tianlin Zhang, Hassan Alhuzali, and Sophia Ananiadou. 2023b. Cluster-level contrastive learning for emotion recognition in conversations. *IEEE Transactions on Affective Computing*, 14:3269–3280.

Kailai Yang, Tianlin Zhang, Shaoxiong Ji, and Sophia Ananiadou. 2023c. A bipartite graph is all we need for enhancing emotional reasoning with commonsense knowledge. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pages 2917–2927.

Seongjun Yun, Minbyul Jeong, Raehyun Kim, Jaewoo Kang, and Hyunwoo J Kim. 2019. Graph transformer networks. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, pages 11983–11993.

Dong Zhang, Weisheng Zhang, Shoushan Li, Qiaoming Zhu, and Guodong Zhou. 2020. Modeling both intra-and inter-modal influence for real-time emotion detection in conversations. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 503–511.

Xiaoheng Zhang and Yang Li. 2023. A cross-modality context fusion and semantic refinement network for emotion recognition in conversation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13099–13110.

Weixiang Zhao, Yanyan Zhao, and Xin Lu. 2022. Cauain: Causal aware interaction network for emotion recognition in conversations. In *Proceedings of the 31st International Joint Conference on Artificial Intelligence*, pages 4524–4530.

Weixiang Zhao, Yanyan Zhao, Xin Lu, Shilong Wang, Yanpeng Tong, and Bing Qin. 2023. Is chatgpt equipped with emotional dialogue capabilities? *arXiv preprint arXiv:2304.09582*.

Wenjie Zheng, Jianfei Yu, Rui Xia, and Shijin Wang. 2023. A facial expression-aware multimodal multi-task learning framework for emotion recognition in

multi-party conversations. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15445–15459.

Peixiang Zhong and Chunyan Miao Di Wang. 2019. Knowledge-enriched transformer for emotion detection in textual conversations. In *EMNLP-IJCNLP*.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*.

Ling Yu Zhu, Zhengkun Zhang, Jun Wang, Hongbin Wang, Haiying Wu, and Zhenglu Yang. 2022. Multi-party empathetic dialogue generation: A new task for dialog systems. In *ACL*, pages 298–307.

Jie Zou, Evangelos Kanoulas, Pengjie Ren, Zhaochun Ren, Aixin Sun, and Cheng Long. 2022. Improving conversational recommender systems via transformer-based sequential modelling. In *Proceedings of the 45th international ACM SIGIR conference on research and development in information retrieval*, pages 2319–2324.

A Feature Extraction

Textual Features: To obtain context-independent utterance-level feature vectors, we follow (Ghosal et al., 2020a) to fine-tune the Roberta Large model to predict emotion labels of utterances. Let an utterance $\mathbf{u}_i$ be a sequence of tokens after applying Byte Pair Encoding (BPE), denoted as $\mathbf{w}_1, \mathbf{w}_2, ..., \mathbf{w}_N$. The emotion label associated with $\mathbf{u}_i$ is represented by $\mathbf{e}_i$, where $\mathbf{e}_i$ belongs to the set of emotion labels E. To prepare the input sequence for the RoBERTa model, we add a special token $[CLS]$ at the beginning of the original utterance. The sequence now becomes $[CLS], \mathbf{w}_1, \mathbf{w}_2, ..., \mathbf{w}_N$. This modified sequence is then fed into the Roberta model. The output of the last layer corresponding to the $[CLS]$ token is used as input to a small feed-forward network, which performs the classification into the appropriate emotion class. After fine-tuning the model for emotion classification, we utilize the model for generating feature vectors. We pass the BPE tokenized utterance appended with $[CLS]$ through the model and extract the outputs from the last four layers corresponding to the $[CLS]$ tokens. The four vectors are combined through averaging, resulting in a context-independent utterance-level feature vector with a dimensionality of 1024.

Acoustic Features: Regarding the extraction of acoustic features, we adopt the methodology outlined in the study by (Majumder et al., 2019), utilizing the openSMILE toolkit (Eyben et al., 2010) for this purpose. The chosen acoustic features undergo a subsequent normalization process, after which a fully connected layer is utilized to achieve dimensionality reduction. Notably, the dimensions of the resulting acoustic features differ between datasets: 1582 for IEMOCAP and reduced to 300 for MELD.

Visual Features: For visual facial expression features, we employ a DenseNet architecture (Huang et al., 2017), which is pre-trained on the Facial Expression Recognition Plus (FER+) corpus (Barsoum et al., 2016), similar to the approach used in (Hu et al., 2021b). The dimension of the visual facial expression feature is 342 for each dataset.

Event		Description
Textual if-then CSK	oEffect	The effect the event has on others besides Person X
	oReact	The reaction of others besides Person X to the event
	oWant	What others besides Person X may want to do after the event
	xAttr	How Person X might be described given their part in the event
	xEffect	The effect that the event would have on Person X
	xIntent	The reason why X would cause the event
	xNeed	What Person X might need to do before the event
	xReact	The reaction that Person X would have to the event
	xWant	What Person X may want to do after the event
Visual if-then CSK	Before	A set of commonsense inferences on events before
	After	A set of commonsense inferences on events after
	Intent	A set of commonsense inferences on people's intents at present.

Table 5: Definitions of the relations of multiple if-then CSK.