Causal-ERC: A Multimodal Framework with Causal Prompting for Emotion Recognition in Conversations with Large Language Models

Ran Jing[1*], **Geng Tu**[1*], **Yice Zhang**[1], **and Ruifeng Xu**[1, 2, 3†]
[1]Harbin Institute of Technology, Shenzhen
[2]Peng Cheng Laboratory
[3]Guangdong Provincial Key Laboratory of Novel Security Intelligence Technologies
{24S051023, 22b951011}@stu.hit.edu.cn, zhangyc_hit@163.com, xuruifeng@hit.edu.cn

Abstract

The rapid advancement of large language models (LLMs) has revitalised research in Emotion Recognition in Conversation (ERC). However, existing LLM-based ERC approaches operate solely on textual input, whereas MLLM-based emotion recognition methods in non-conversational scenarios typically perform only basic multimodal fusion and fail to consider speaker-sensitive contextual dependencies, which limits their performance on ERC tasks. To integrate multimodal cues effectively and address their limitations in handling contextual dependencies, we propose a novel LLM-based framework, Causal-ERC, which captures context representations within each modality and incorporates them into the LLM. Moreover, experimental results show that LLMs perform poorly on long conversations. To improve LLMs' ability to model long conversations, we adjust corresponding causal prompts according to the causal type of each utterance. Experiments on two benchmark MERC datasets demonstrate that our Causal-ERC framework consistently outperforms existing state-of-the-art approaches and improves LLM's performance in long-context scenarios.

1 Introduction

Emotion recognition in conversations (ERC) aims to identify the emotion of each utterance in conversations. This task plays a crucial role in applications such as recommendation systems and dialogue generation (Tu et al. 2022). As the development of ERC, researchers found that in real-world applications, conversations often contain not only textual information but also audio and visual modalities. Fusing and leveraging information from different modalities can significantly improve the accuracy of emotion prediction, cause they conclude the tone and facial expressions of speakers (Li et al. 2022). As a result, multimodal fusion has attracted many researchers in the ERC. In recent years, with the rise of large language models (LLMs), their strong capabilities in contextual understanding and reasoning have been increasingly leveraged in the text-only ERC task.

Despite advances in LLM-based ERC, they still face two key challenges: (1) **Existing LLM-based ERC methods lack consideration of multimodal information.** Despite multimodal information being important in ERC, existing LLM-based ERC approaches, such as InstructERC (Lei et al. 2023), operate solely on textual input. While MLLM-based emotion recognition methods in non-conversational scenarios, like Emotion-LLaMA (Cheng et al. 2024) and AffectGPT (Lian et al. 2024), did not consider conversational structure and speaker-aware contextual dependencies. (2) **LLMs struggle to model long-range contextual dependencies.** Experimental results reveal that LLMs perform poorly on long conversations, which may be due to their neglect of two underlying causal relationships: when utterances drive emotions, we name this C1; and when emotions drive utterances, which is named C2. These two directions are closely aligned with the dual system theory in psychology, which distinguishes between fast and slow thinking (Kahneman 2011). Specifically, emotion can affect the use of language (Barrett 2006), indicating a more reactive,

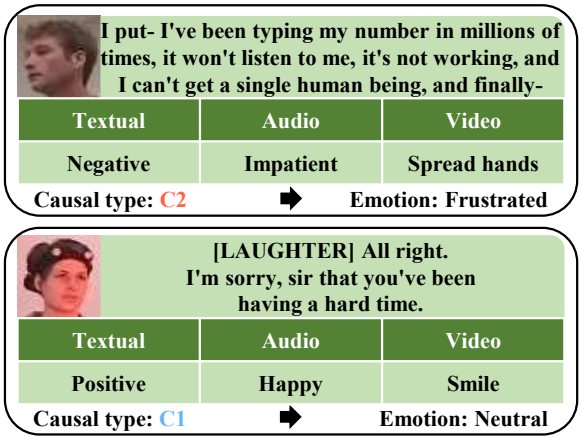

Figure 1: Samples that illustrate the role of causal prompting in ERC. In the first example, both the utterance and facial expression suggest anger, but causal analysis reveals the emotion stems from disgust toward the AI customer service. In the second example, the utterance and facial expression appear to convey happiness. However, contextual information reveals that the speaker is a customer service agent, and the content is work-related, indicating a neutral emotion.

*These authors contributed equally.
†Corresponding authors.

fast-thinking mode. In contrast, language can be the cause of emotion(Satpute et al. 2013), reflecting a deliberative, slow-thinking process on the part of the speaker. Previous studies (Tu et al. 2023) in ERC have indirectly supported this perspective by observing that some utterances' emotions can be accurately identified without context, while others rely on rich contextual information (Tu et al. 2025). As shown in Fig. 1, properly analysing causal relationships significantly improves the ability of LLM in ERC.

To solve the above problems, we propose a novel MLLM-based framework with causal prompting for MERC, Causal-ERC, that can fuse multimodal information and improve the ability of LLMs on modelling long-term context. Specifically, to model the context with speaker information, we propose a DialogueRNN (Majumder et al. 2019) layer on each modality and extract the characteristics. Then, a multimodal fusion model based on bidirectional multihead cross-attention layers is used to integrate multimodal information. The fused multimodal information will be fed to the LLM with a conversational prompt. Moreover, to enhance the LLM's ability in long-term context modelling, we analyse the causal relationship of each utterance by leveraging the Peak-End Rule (Epstein 1994), which suggests that people judge experiences based on the most intense ("peak") and final ("end") moments rather than the average. If a target utterance's emotional intensity aligns more with the average intensity of prior utterances, it indicates slow thinking (Causal Type C1). Conversely, if it aligns with the average of the peak and end intensities, it reflects fast thinking (Causal Type C2). Based on the causal relationship, we design a corresponding causal prompt for the LLM.

In summary, our contributions are as follows:

- We are the first to use LLM to reason with multimodal information on the MERC task.
- We identify causal relationships and generate causal prompts to guide LLMs in context modelling.
- Experimental results on two popular datasets show that our proposed Causal-ERC framework outperforms state-of-the-art baselines and improves the LLM's performance on long-context scenarios.

2 Related Work

MERC: The emotion generation theory (Gross and Barrett 2011) emphasizes the importance of contextual information for identifying emotions. Early studies focused on textual modality and modelled context and speaker information (Majumder et al. 2019; Ghosal et al. 2019). With the increasing relevance of multi-modality in real-world applications, MultiModal Emotion Recognition in Conversation (MERC) has become a focal point of research (Shi and Huang 2023). In MERC, the fusion of different modalities is crucial, with various approaches being explored, from aggregation-based methods (Hazarika et al. 2018; Lian, Liu, and Tao 2021) to graph-based fusion approaches (Chen et al. 2023; Nguyen et al. 2023; Tu et al. 2024). In recent years, LLMs have achieved remarkable success across a wide range of domains. Therefore, some researchers have applied them to ERC tasks and achieved promising performance (Lei et al. 2023; Xue et al. 2024). However, LLMs still face challenges in modelling long-range contexts.

Cause-Effect Discovery: Traditional methods for causal discovery, such as constraint-based and score-based approaches, have limited direct applicability in NLP due to the unstructured and high-dimensional nature of language data. Instead, researchers have increasingly adopted representation learning and neural methods to uncover latent causal structures from text. Causal reasoning has progressed from early rule-based methods (Girju, Badulescu, and Moldovan 2006; Mirza and Tonelli 2014) to approaches leveraging pre-trained language models for capturing implicit and contextual causality (Sap et al. 2019; Rashkin et al. 2018). Neural causal discovery methods such as CGNN (Goudet et al. 2018) and NOTEARS (Zheng et al. 2018) have inspired adaptations that infer causal structures over discourse elements. Recently, LLMs have shown promising capabilities in encoding causal knowledge (Jin et al. 2023; Wu et al. 2024), paving the way for more explainable and reasoning-aware NLP systems. Despite its potential, cause-and-effect reasoning has not yet been explored in ERC tasks.

3 Methodology

3.1 Task Definition

Let $\mathbf{U} = [\mathbf{u}_1, ..., \mathbf{u}_\mathbf{N}]$ be a conversation with $\mathbf{N}$ utterances from $\mathbf{M} \geq 2$ speakers, where each $\mathbf{u}_i$ is spoken by $\mathbf{sp}_i$. The features of each utterance $\mathbf{u}_i$ is represented by a triplet $\mathbf{x}_i = \{\mathbf{x}_i^a, \mathbf{x}_i^v, \mathbf{x}_i^t\}$. $\mathbf{x}_i^a \in \mathbb{R}^{d_a}$, $\mathbf{x}_i^v \in \mathbb{R}^{d_v}$, and $\mathbf{x}_i^t \in \mathbb{R}^{d_t}$ denote the acoustic, visual, and textual features of $\mathbf{u}_i$, respectively. ERC aims to predict the emotion label $\widehat{\mathbf{y}}_i$ for each utterance. The set of predefined emotions is $\mathbf{Y} = [\mathbf{y}_1, ..., \mathbf{y}_\kappa]$, where κ is the number of emotion categories.

3.2 Framework Overview

To assist the LLM in modelling contextual information more effectively, we propose Causal-ERC, which uses an LLM to analyse multimodal information. As illustrated in Fig. 2, Causal-ERC transforms utterances into sequences through ERC prompting, which can be input into the LLM to perform reasoning tasks. Meanwhile, Causal-ERC models the contextual information with various modalities, fuses them effectively, and then feeds the fused representation together with the sequence gained by ERC prompting into the LLM for emotion recognition. Moreover, during each iteration, Causal-ERC analyses the emotional intensity of each utterance to categorise causal relationships. In the next iteration, Causal-ERC selects the corresponding causal prompt to assist the LLM in modelling long-term context.

3.3 ERC Prompting

Following InstructERC, to effectively adapt LLMs to the ERC task, we design our prompts with three key components: an instruction, the context, and a label statement.

Instruction: The instruction component explicitly defines the role. So $\mathcal{P}_{ins}$ is defined as follows:

```
Now you are an expert in sentiment
and emotional analysis. The following
```

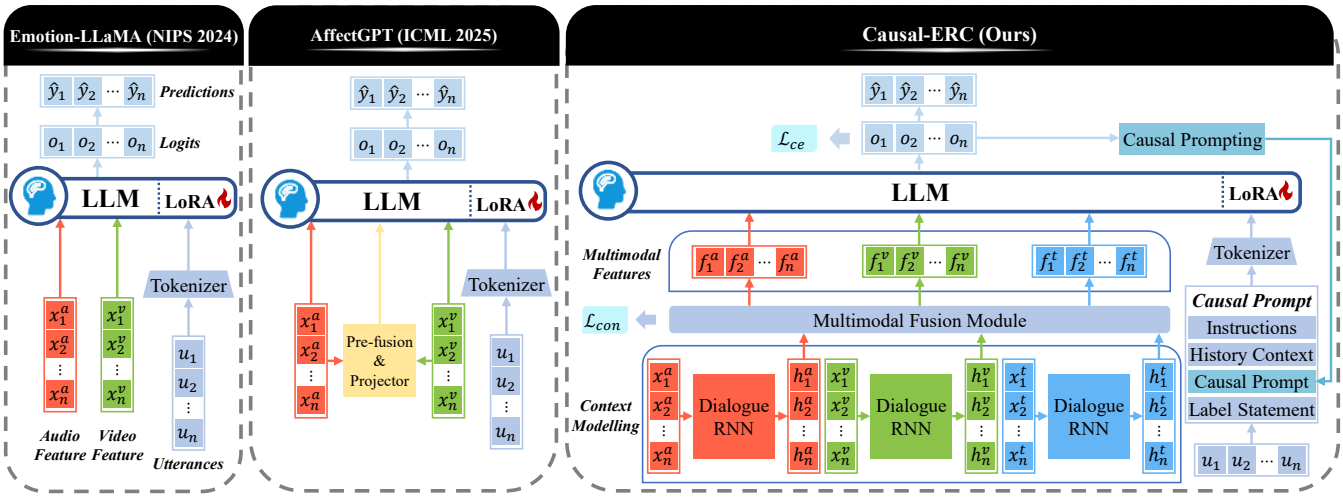

Figure 2: A structural comparison diagram of Emotion-LLaMA, AffectGPT, and Causal-ERC. Compared to the previous methods, our proposed Causal-ERC introduces speaker-aware contextual fusion and multimodal integration. In addition, we design a causal prompting mechanism to enhance the LLM's ability to model long-range dependencies in conversations.

`is the conversation text which involves several speakers.`

History Context: To preserve information from previous and current utterances, we incorporate utterances along with their respective speakers. For the target utterance $\mathbf{u}_i$, its context prompt $\mathcal{P}_{his}$ consists of all preceding and current utterances, which is formatted as follows:

$\mathbf{sp}_1 : \mathbf{u}_1 +$ <TAB> $+ \ldots + \mathbf{sp}_i : \mathbf{u}_i$

Label Statement: To restrict the model's output to a predefined set of emotions and guide its focus on the current utterance, we construct the label statement $\mathcal{P}_{lab}$ as follows:

`Please select the emotional label of` $<\mathbf{sp}_i : \mathbf{u}_i>$ `from` $<\mathbf{y}_1, \ldots, \mathbf{y}_\kappa>$

By combining the three components mentioned above, we transform an utterance $\mathbf{u}_i$ within a conversation into a sequence format $\mathbf{s}_i$, resulting in the final prompt: $\mathcal{P} = \mathcal{P}_{ins} + \mathcal{P}_{his} + \mathcal{P}_{lab}$. Then, we input $\mathbf{s}_i$ into an LLM-based model f_θ, which generates a prediction vector $\mathbf{o}_i = [\mathbf{o}_{i1}, ..., \mathbf{o}_{i\kappa}]$. Here, $\mathbf{o}_{ij}$ represents the probability that f_θ predicts the emotional category of utterance $\mathbf{u}_i$ as $\mathbf{y}_j$.

3.4 Multimodal Fusion

To effectively utilise multimodal information, we first model the features of each modality, fuse them into a unified representation, and then feed the result into the LLM.

Context Modelling: To model speaker-aware context information, we utilise DialogueRNN as the encoder to derive a hidden representation $\mathbf{h}_i$ for each utterance:

$$\mathbf{SP} = [\mathbf{sp}_1, \mathbf{sp}_2, ..., \mathbf{sp}_n] \tag{1}$$

$$[\mathbf{h}_1^\xi, \mathbf{h}_2^\xi, ..., \mathbf{h}_n^\xi] = \mathbf{E}([\mathbf{x}_1^\xi, \mathbf{x}_2^\xi, ..., \mathbf{x}_n^\xi], \mathbf{SP}) \tag{2}$$

where $\mathbf{SP}$ denotes the speaker information of the whole conversation, $\xi \in \{a, v, t\}$ represents audio, video, and textual modality, respectively. $\mathbf{h}_i^\xi$ is the hidden representation with speaker-aware context information of $\mathbf{u}_i$, $\mathbf{E}(.)$ represents the DialogueRNN encoder.

Multimodal Fusion: To effectively integrate multimodal information and capture consistency among different modals, we design a model to perform multimodal fusion, which consists of three multi-head attention layers: $\mathbf{Attn}_a$, $\mathbf{Attn}_v$, and $\mathbf{Attn}_t$.

$$\mathbf{f}_i^a = \mathbf{Attn}_a(\mathbf{h}_a, \mathbf{h}_v, \mathbf{h}_t) \tag{3}$$

$$\mathbf{f}_i^v = \mathbf{Attn}_v(\mathbf{h}_v, \mathbf{h}_t, \mathbf{h}_a) \tag{4}$$

$$\mathbf{f}_i^t = \mathbf{Attn}_t(\mathbf{h}_t, \mathbf{h}_a, \mathbf{h}_v) \tag{5}$$

where $\mathbf{f}_i^\xi$ indicates the fused feature of each modality, they have the same dimensional space.

Consistency Loss: We utilise a Soft Hirschfeld Gebelein-Rényi loss (Wang et al. 2019) to maximise the correlations across multimodal-fused textual, audio and visual features extracted from the multimodal fusion module. The Soft-HGR loss is defined as the following equations:

$$\mathcal{L}_{con} = -\sum_{\mathbf{Q}\neq\mathbf{V},\mathbf{Q},\mathbf{V}\in\mathbf{F}} (\mathbb{E}[\mathbf{Q}^T\mathbf{V}] - \frac{1}{2}tr(cov(\mathbf{Q})cov(\mathbf{V})))$$
$$s.t. \mathbb{E}[\mathbf{Q}] = 0, \forall \mathbf{Q} \in \mathbf{F} \tag{6}$$

where $\mathbf{F} = \{\mathbf{F}^t, \mathbf{F}^a, \mathbf{F}^v\}$, $\mathbf{F}^t = [\mathbf{f}_1^t, \mathbf{f}_2^t, ..., \mathbf{f}_n^t]^T$, $\mathbf{F}^a = [\mathbf{f}_1^a, \mathbf{f}_2^a, ..., \mathbf{f}_n^a]^T$, $\mathbf{F}^v = [\mathbf{f}_1^v, \mathbf{f}_2^v, ..., \mathbf{f}_n^v]^T$. Expectations and covariances are approximated through sample means and sample covariances.

Multimodal Integrating with Prompts: To integrate multimodal features with prompts, following (Cheng et al. 2024), we introduce trainable linear mappings σ^ξ to convert multimodal feature $\mathbf{f}_i^\xi$ into language embedding tokens $\mathcal{T}$:

$$\mathcal{T}^\xi = \sigma^\xi \cdot \mathbf{f}_i^\xi, \xi \in \{a, v, t\} \tag{7}$$

Where $\mathcal{T}^\xi$ stands for the resulting multimodal tokens of each modal feature, these tokens are fused with prompt tokens through the inner cross-attention mechanism of Causal-ERC, enabling it to capture and reason about the emotional content in the multimodal input.

Algorithm 1: The process of Causal Prompting

Input : $\mathbf{U} = [\mathbf{u}_1, ..., \mathbf{u}_\mathbf{N}]$, LLM-based model f_θ
Output: Predicted emotional labels $\widehat{\mathbf{Y}} = [\widehat{\mathbf{y}}_1, ..., \widehat{\mathbf{y}}_\mathbf{N}]$
# Calculate emotional intensities
for $i \leftarrow 1$ **to N do**
  # Original prompt prediction
  $\mathbf{o}_i \leftarrow f_\theta(\mathcal{P}, \mathbf{u}_i)$
  $\mathcal{I}_i \leftarrow \|\mathbf{o}_i\|_2$
for $e \leftarrow 1$ **to max_epoch do**
  for $i \leftarrow 1$ **to N do**
    $\mathcal{C}_{Mean} \leftarrow mean(\mathbf{o}_1, ..., \mathbf{o}_{i-1})$
    # Select the peak utterance
    $\mathbf{o}_{Peak} \leftarrow \arg\max_{\mathbf{o} \in \{\mathbf{o}_1, ..., \mathbf{o}_{i-1}\}}(\|\mathbf{o}\|_2)$
    $\mathbf{o}_{End} \leftarrow \mathbf{o}_{i-1}$
    $\lambda_{C1} \leftarrow \|\mathcal{I}_i - \mathcal{C}_{Mean}\|_2$
    $\lambda_{C2} \leftarrow \|\mathcal{I}_i - \frac{1}{2}(\mathbf{o}_{Peak} + \mathbf{o}_{End})\|_2$
    if $\lambda_{C1} < \lambda_{C2}$ **then**
      $\mathcal{P}_{cau} \leftarrow \mathcal{P}_{C1}$
    else
      $\mathcal{P}_{cau} \leftarrow \mathcal{P}_{C2}$
    $\mathcal{P}' \leftarrow \mathcal{P}_{ins} + \mathcal{P}_{his} + \mathcal{P}_{cau} + \mathcal{P}_{lab}$
    $\mathbf{o}'_i \leftarrow f_\theta(\mathcal{P}', \mathbf{u}_i)$ // Final prediction
    # Select highest-probability category as result
    $\widehat{\mathbf{y}}_i \leftarrow \arg\max_j(\mathbf{o}'_{ij})$
  for $i \leftarrow 1$ **to N do**
    $\mathbf{o}_i \leftarrow \mathbf{o}'_i$ // prepare logits for the next iteration
return $\widehat{\mathbf{Y}}$

3.5 Causal Prompting

To enhance the LLM's ability to model long-term context, we introduce a causal prompting strategy. Firstly, we apply the Peak-End Rule to classify the causal relationship of each utterance based on the emotional intensity. Then, we design tailored causal prompts according to the identified causal types. The overall procedure is outlined in Algorithm 1.

Causality Categorising: We categorise the causal relationship in a conversation into two distinct types according to previous work. In causal type C1, language can be the cause of emotion, where emotion serves as the effect, while in causal type C2, emotion can affect the use of language, where emotion serves as the cause.

Vectorized Intensity: To track emotional fluctuations in the conversation, we capture the emotional intensity of each utterance. We define the emotional intensity $\mathcal{I}_i$ as the L2 norm of the logits vector $\mathbf{o}_i$. Moreover, we evaluate emotional shifts between utterances with the Euclidean distance between vectors $\|\mathbf{o}_i - \mathbf{o}_j\|_2$.

Peak-End Rule: The peak-end rule is a psychological heuristic suggesting that people judge an experience largely based on how they felt at its most intense moment (the "peak") and at its end, rather than by the average of every moment of the experience. This rule explains why individuals' memories of events are often shaped disproportionately by emotionally charged moments and final impressions. In the context, this means that certain utterances, especially those with high emotional intensity or those occurring at the end, can have a stronger impact on how the overall conversation is perceived emotionally.

- **Causal Type C1:** In C1, emotional responses are more deliberate. The speaker carefully considers details and analyses rationally. As a result, $\mathbf{o}_i$ tends to align with $\mathcal{C}_{Mean} = 1/\mathbf{N}\sum_{i=1}^{\mathbf{N}} \mathbf{o}_i$. Therefore, $\lambda_{C1} = \|\mathcal{C}_{Mean} - \mathbf{o}_i\|_2$ indicates the proximity of the current utterance to the causal relationship C1. A smaller value of λ_{C1} indicates that the target utterance is more likely to exhibit the causal relationship C1.
- **Causal Type C2:** In C2, emotional responses are quick and intense. The speaker feels emotions first and expresses them verbally afterwards. Thus, $\mathbf{o}_i$ is likely influenced by $\mathbf{o}_{peak}$ (the maximum emotion intensity) and $\mathbf{o}_{end} = \mathbf{o}_{i-1}$ (the last utterance in history). As a result, $\lambda_{C2} = \|(\mathbf{o}_{Peak} + \mathbf{o}_{End})/2 - \mathbf{o}_i\|_2$ indicates the proximity of the current utterance to the causal relationship C2. A smaller λ_{C2} indicates that the target utterance is more likely to exhibit the causal relationship C2.

Based on the above, if $\lambda_{C1} < \lambda_{C2}$, it indicates that the causal relationship is more likely to be C1; otherwise, it is more likely to be C2. Specifically, for overly long conversations, we set a history window size N_ω and select only the last N_ω utterances to prevent distant context from interfering with the judgment of causal relationships.

Prompt Design: We design a specific prompt for LLM according to the causal relationship. For C1, we supply the LLM with longer context. If the length of the history context is l in the baseline, it will be increased to $l * \lambda_{C2}/\lambda_{C1}$. Moreover, the causal prompt $\mathcal{P}_{cau}$ will be set to $\mathcal{P}_{C1}$:

`In this utterance, history utterances cause the emotion of` $\mathbf{sp}_i$`. As a result, he speaks the current utterance. So it is important to consider historical utterances.`

For C2, we guide the LLM to focus on the emotion of the peak utterance, the causal prompt $\mathcal{P}_{cau}$ will be set to $\mathcal{P}_{C2}$:

`In this utterance,` $\mathbf{sp}_i$ `first experiences his emotion, which then leads to the production of the current utterance. So it is necessary to focus on the <` $\mathbf{u}_{Peak}$ `> to infer the underlying emotion.`

Here, $\mathbf{u}_{Peak} = \arg\max_{\mathbf{u}_i}(\|\mathbf{o}_i\|_2)$ represents the utterance with the most intense emotional activation.

During each iteration, the LLM-based model f_θ processes the causal prompting $\mathcal{P}' = \mathcal{P}_{ins} + \mathcal{P}_{his} + \mathcal{P}_{cau} + \mathcal{P}_{lab}$. Then $\mathcal{P}'$ is used to convert conversations to sequences, which serve as input to the LLM.

Dynamic Prompting: Based on the prediction of emotions from the previous epoch, the model determines the causal relationship and subsequently adjusts the corresponding causal prompts. This updated prompt is then used as input for the next training epoch, allowing the model to adapt its reasoning strategy dynamically. Following each iteration, the model updates its parameters, evolving into $f_{\theta'}$, which

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.

serves as the new baseline for subsequent training. In this dynamic prompt framework, on the one hand, causal prompts enhance the model's ability for contextual modelling and emotional reasoning; on the other hand, the improved accuracy of emotion prediction conversely enables a more reliable classification of causal relationships.

4 Experiments

4.1 Datasets

We conduct experiments on the IEMOCAP (Busso et al. 2008) and MELD (Poria et al. 2018) datasets as they are the most widely used and representative benchmarks for MERC. The statistics are presented in Table 1.

IEMOCAP consists of sessions where actors perform scripted scenarios. Each utterance is labelled with one of the emotions: happy, angry, neutral, sad, excited, or frustrated.

MELD is a multi-party conversation dataset collected from the TV show *Friends*, which is an extension of the EmotionLines dataset (Chen et al. 2018). Each utterance is annotated with one of the emotions: surprise, fear, disgust, anger, sadness, neutral, or joy.

4.2 Comparison Models

We benchmark our framework against a range of ERC models:

Smaller Models: We compare our proposed Causal-ERC with various MERC methods, including recurrent-based networks (DialogueRNN (Majumder et al. 2019)), transformer-based networks (MultiEmo (Shi and Huang 2023), CMCF-SRNet (Zhang and Li 2023), SDT (Ma et al. 2023), TelME (Yun et al. 2024)), and graph-based networks (MMDFN (Hu et al. 2022), CORECT (Nguyen et al. 2023), M3Net (Chen et al. 2023)).

LLM-based Models: InstructERC (Lei et al. 2023) reformulates the ERC task from a discriminative framework to a generative framework. BiosERC (Xue et al. 2024) extracts the "biographical information" of the speaker as supplementary knowledge. Moreover, we prompt ChatGPT on the MERC task and select pictures from videos as visual input.

4.3 Experimenal Settings

Because no existing MLLM has yet been applied to the MERC task, we modify Emotion-LLaMA and apply it to MERC as the baseline. All experiments are conducted on CUDA version 11.7. We select Llama-3.1-8B-Instruct (Dubey et al. 2024) as the basic LLM. We conduct a hyperparameter search for our proposed Causal-ERC on each dataset by hold-out validation with a validation set. The hyperparameters to search include learning rate, batch size and epoch. Given the efficiency and effectiveness of Parameter-Efficient Fine-Tuning (PEFT), we adopt LoRA (Hu et al. 2021) and insert low-rank adapters after the self-attention layers. The adapter dimension is set to 16. All reported results are averaged from 5 random test set runs.

4.4 Overall Results

Table 2 compares our proposed Causal-ERC with other methods. For early aggregation-based fusion, as seen in models like DialogueRNN and MMDFN, they often overlook the complex interactions between modalities, resulting in suboptimal utilisation of contextual cues. Therefore, most current MERC methods, such as M^3Net and MultiEMO, employ graph-based fusion methods to capture interactions between modalities, achieving enhanced performance. In recent years, LLMs have started to be popular in ERC due to their strong capabilities in contextual understanding. While BiosERC outperforms all previous works, zero-shot ChatGPT even gets better performance than some early aggregation-based models.

Nonetheless, these methods typically lack consideration for the underlying causal dynamics between emotion and context, and LLM-based methods cannot effectively leverage multimodal information. Causal-ERC dynamically determines the model's focus by leveraging the causal relationship between emotion and context, allowing it to selectively attend to the most informative content. This approach achieves state-of-the-art performance in W-F1 scores compared to previous methods, setting a new benchmark in the field. The removal of Causal-ERC causes a decrease of 2.83% in the F1 score for the IEMOCAP dataset and 2.41% for the MELD dataset. This difference in performance improvement may be attributed to the fact that conversations in the IEMOCAP dataset are generally longer and contain more information compared to those in MELD. Moreover, our proposed Causal-ERC achieves state-of-the-art performance in W-F1 scores compared to previous methods, which proves the effectiveness of this framework.

4.5 Ablation Study

To investigate the impact of each component of Causal-ERC, we conducted an ablation study, whose results are shown in Table 3. **w/o** represents the removal operation. The results suggest that all components have worked and all the improvements are statistically significant, as evidenced by the paired t-test results with a p-value < 0.05.

Analysis of Multimodal Fusion: We implement the Multimodal Fusion module to capture multimodal information and use $\mathcal{L}_{con}$ to improve the quality of them. In Table 3, the removal of $\mathcal{L}_{con}$ results in a slight decrease on the IEMOCAP and MELD datasets. However, the removal of the whole Multimodal Fusion Module leads to a large decrease of 1.75% on the IEMOCAP dataset and 1.87% on the MELD dataset. To further prove that Causal-ERC can effectively leverage multimodal information through multimodal fusion, we compare the performance of Causal-ERC on different modality settings. Since Causal-ERC relies on textual modality to provide prompts for the LLM, we only conduct modality comparisons including the textual modality. As shown in Table 4, the performance of Causal-ERC

Patterns	Methods	IEMOCAP							MELD							
		Happy	Sad	Neutral	Angry	Excited	Frustrated	W-F1	Neutral	Surprise	Fear	Sadness	Joy	Disgust	Anger	W-F1
AVT	★DialogueRNN$^\flat$	32.20	80.26	57.89	62.82	73.87	59.76	62.89	76.97	47.69	-	20.41	50.92	-	45.52	57.66
AVT	★MMDFN$^\sharp$	42.22	78.98	66.42	69.77	75.56	66.33	68.18	77.76	50.69	-	22.93	54.78	-	47.82	59.46
AVT	★M^{3}Net$^\flat$	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
AVT	★MultiEMO$^\natural$	52.46	83.44	71.46	66.26	73.86	67.77	70.61	79.05	56.75	20.78	41.07	64.72	29.36	53.48	65.63
AVT	CMCF-SRNet$^\natural$	52.20	80.90	68.80	**70.30**	76.70	61.60	69.60	-	-	-	-	-	-	-	62.30
AVT	★CORECT$^\natural$	59.30	80.53	66.94	69.59	72.69	68.50	70.02	-	-	-	-	-	-	-	-
AVT	★ SDT‡	56.21	75.66	68.64	65.62	**80.28**	63.36	69.19	79.65	58.10	7.55	42.58	63.65	14.46	51.74	65.14
AVT	★ TelME$^\natural$	-	-	-	-	-	-	70.48	80.22	60.33	26.97	43.45	65.67	26.42	56.70	67.37
VT	ChatGPT‡	52.08	64.48	54.19	37.67	39.47	52.70	50.76	73.88	50.61	**33.33**	41.42	54.20	38.41	53.76	61.39
T	★ BiosERC$^\natural$	-	-	-	-	-	-	71.19	-	-	-	-	-	-	-	69.83
T	★ InstructERC$^\natural$	-	-	-	-	-	-	71.39	-	-	-	-	-	-	-	69.15
AVT	Baseline‡	58.90	**81.18**	69.22	66.28	64.16	70.01	69.06	80.34	57.79	27.91	44.38	65.20	44.93	58.05	67.84
AVT	Causal-ERC (Ours)	**62.95**	80.61	**72.80**	67.76	74.14	**68.83**	**71.89**	**81.84**	**62.96**	31.75	**47.67**	**67.69**	**45.53**	**61.03**	**70.25**

Table 2: Per-class F1 scores (%) and W-F1 (%) comparison between our method and different ERC models in different modality settings (A, V, T represent acoustic, visual, and textual modalities, respectively). ★ indicates the source code has been released. ‡ denotes our re-implementation results. $^\sharp$, $^\flat$, and $^\natural$ represent results from (Hu et al. 2022), (Shi et al. 2023), and original papers.

Methods	IEMOCAP							MELD						
	Happy	Sad	Neutral	Angry	Excited	Frustrated	W-F1	Neutral	Surprise	Fear	Sadness	Joy	Disgust	Anger	W-F1
Ours	62.95	80.61	72.80	67.76	74.14	68.83	71.89	81.84	62.96	31.75	47.67	67.69	45.53	61.03	70.25
w/o $\mathcal{L}_{con}$	55.12	85.00	74.78	66.49	74.10	63.82	71.01	81.39	62.84	32.00	42.21	67.71	41.44	60.19	69.37
w/o Multimodal Fusion	61.39	82.97	70.93	64.39	77.05	61.56	70.14	80.00	59.11	33.66	46.29	66.67	43.06	58.97	68.38
w/o Causal Prompting	54.92	81.65	69.88	67.02	76.18	67.70	70.68	80.99	64.93	30.77	48.69	67.96	51.09	49.81	68.82

Table 3: Ablation results of Causal-ERC, which shows the impact of key components on W-F1 performance (%).

Patterns	IEMOCAP		MELD	
	Acc	WF1	Acc	WF1
T	69.19	69.13	69.43	68.10
AT	70.12	70.18	69.62	68.80
VT	71.29	71.23	69.81	69.50
AVT	71.66	71.89	70.50	70.25

Table 4: Accuracy and W-F1 performance (%) across different modalities. A, V, and T represent the acoustic, visual, and textual modalities, respectively.

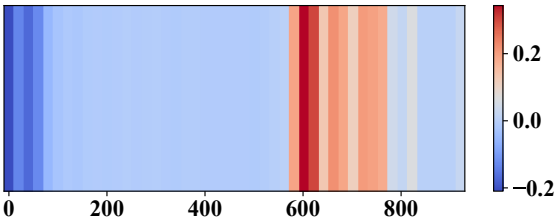

Figure 3: The improvement in attention scores of Causal-ERC compared to Causal-ERC w/o Causal Prompting.

improves accordingly with the increase of modality number. This indicates the ability of Causal-ERC to integrate and utilise multimodal information effectively.

Analysis of Causal Prompting: To improve the LLM's ability in long-term context modelling, we introduce causal prompting on Causal-ERC. In Fig. 3, we compare the attention distributions of Causal-ERC and Causal-ERC w/o Causal Prompting at different token positions. The results show that after introducing causal prompting, the LLM assigns higher attention scores to more distant tokens, at the cost of reduced attention to nearby tokens. This indicates that causal prompting effectively enhances the LLM's ability to model long-range context.

4.6 Long-term Conversation Analysis

In Fig. 4, we illustrate the performance improvements of Causal-ERC over the baseline model. Causal-ERC consistently outperforms the baseline across various emotion categories and conversation positions. Notably, the baseline model performs relatively poorly in long conversation positions, indicating its difficulty in capturing long-term contextual dependencies. However, Causal-ERC mitigates this weakness to some extent, demonstrating improved performance in longer conversations. As shown in Fig. 5, after introducing causal prompting, Causal-ERC achieves substantial performance improvements, especially on long conversations. This suggests that Causal-ERC enhances the model's ability to retain and utilise contextual information more effectively throughout extended conversations.

4.7 Hyper-parameter Analysis

As shown in Fig. 6, we evaluate the impact of varying the history window size $\mathbf{N}_\omega$ on model performance across different datasets. On IEMOCAP, performance consistently improves with larger history windows and peaks at $\mathbf{N}_\omega = 8$,

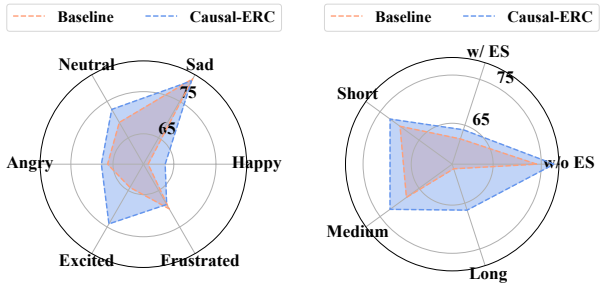

Figure 4: Performance of Causal-ERC and Baseline across different emotional categories, types (ES: Emotion Shift) and conversation positions (Short: first 1/3, Medium: middle 1/3, and Long: last 1/3) on the IEMOCAP dataset.

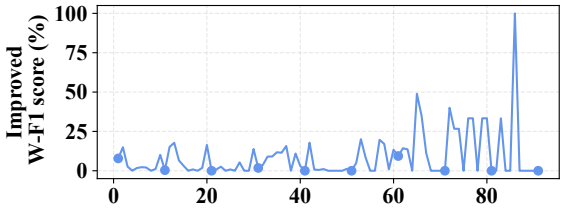

Figure 5: Lifting performance of Causal-ERC on different positions in conversations on the IEMOCAP dataset.

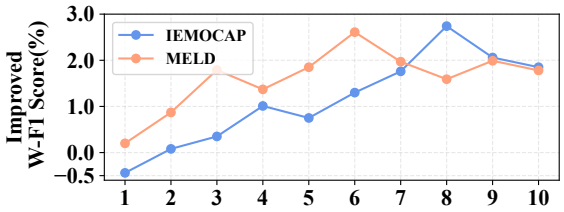

Figure 6: Lifting performance of Causal-ERC compared to the Baseline across different history window sizes on the validation dataset.

after which it begins to decline. A similar trend is observed on MELD, though the performance peaks earlier at $\mathbf{N}_\omega = 6$. This difference can be attributed to the fact that conversations in IEMOCAP are generally longer and more context-dependent than those in MELD.

4.8 Case Study

As shown in Fig. 7, we select a conversation between a male and a female speaker as a case study to demonstrate Causal-ERC's improvement in emotion recognition. For the second-to-last utterance, a surface-level interpretation might lead to the mistaken assumption that the male is angry due to an unsuccessful service request. However, within the Causal-ERC framework, the causal type of this utterance is C2, where the current utterance is emotionally driven. By referencing cues from the peak utterance, Causal-ERC effectively infers that the male's emotion is not anger toward the service itself, but rather disgust toward the intelligent customer service system. Moreover, for the final utterance, the baseline model incorrectly classifies the emotion as happy simply due to the presence of "[laughter]." In contrast, Causal-ERC recognises the causal type of this utterance is C2, the female is likely expressing polite or formal laughter, and thus correctly categorises the overall emotion as neutral.

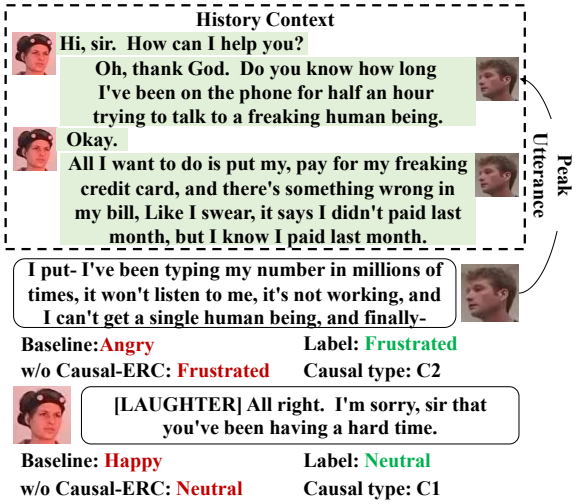

Figure 7: Examples of conversation in the IEMOCAP dataset for the case study. The golden labels for the utterances are highlighted in green font.

4.9 Error Analysis

Many prediction errors of our Causal-ERC framework are related to class imbalance, which is evidenced by the low F1 scores of 31.75% and 45.53% for the 'Fear' and 'Disgust' emotions in the MELD dataset. Additionally, emotional shift remains a persistent challenge. To validate this claim, we analysed the performance of Causal-ERC under both ES and non-ES scenarios. Causal-ERC's performance on conversations with emotion shift samples is consistently lower than that without emotion shift, with a WF1 score drop of 13.63% on the IEMOCAP dataset and 12.71% on the MELD dataset. This may be because the emotional shifts lead the model to capture incorrect causal relationships.

5 Conclusion

In this paper, we propose Causal-ERC, a novel MLLM-based framework with causal prompting for MERC. Causal-ERC fuses multimodal information and conducts context modelling with speaker information. Moreover, Causal-ERC can analyse the causal relationship of each utterance through the Peak-End Rule and selects the corresponding causal prompt to improve the ability of long-term context modelling. Extensive experiments on IEMOCAP and MELD datasets validate the effectiveness of Causal-ERC.

Acknowledgments

We thank all anonymous reviewers for their helpful comments. This work was partially supported by the National Natural Science Foundation of China 625B2060 and 62576120, the Major Key Project of PCL2025A09 and Key Laboratory of Computing Power Network and Information Security, Ministry of Education under Grant No.2024ZD020 and Fundamental Research Funds for the Central Universities, HIT.DZJJ.2025050.

References

Barrett, L. F. 2006. Solving the emotion paradox: Categorization and the experience of emotion. *Personality and social psychology review*, 10(1): 20–46.

Busso, C.; Bulut, M.; Lee, C.-C.; Kazemzadeh, A.; Mower, E.; Kim, S.; Chang, J. N.; Lee, S.; and Narayanan, S. S. 2008. IEMOCAP: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42: 335–359.

Chen, F.; Shao, J.; Zhu, S.; and Shen, H. T. 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*, 10761–10770.

Chen, S.-Y.; Hsu, C.-C.; Kuo, C.-C.; Ku, L.-W.; et al. 2018. Emotionlines: An emotion corpus of multi-party conversations. *arXiv preprint arXiv:1802.08379*.

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

Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; Schelten, A.; Yang, A.; Fan, A.; et al. 2024. The llama 3 herd of models. *arXiv e-prints*, arXiv–2407.

Epstein, S. 1994. Integration of the cognitive and the psychodynamic unconscious. *American psychologist*, 49(8): 709.

Ghosal, D.; Majumder, N.; Poria, S.; Chhaya, N.; and Gelbukh, A. 2019. DialogueGCN: A Graph Convolutional Neural Network for Emotion Recognition in Conversation. In *In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, 154–164.

Girju, R.; Badulescu, A.; and Moldovan, D. 2006. Automatic discovery of part-whole relations. *Computational Linguistics*, 32(1): 83–135.

Goudet, O.; Kalainathan, D.; Caillou, P.; Guyon, I.; Lopez-Paz, D.; and Sebag, M. 2018. Learning functional causal models with generative neural networks. *Explainable and interpretable models in computer vision and machine learning*, 39–80.

Gross, J. J.; and Barrett, L. F. 2011. Emotion Generation and Emotion Regulation: One or Two Depends on Your Point of View. *Emotion Review*, 3(1).

Hazarika, D.; Poria, S.; Mihalcea, R.; Cambria, E.; and Zimmermann, R. 2018. Icon: Interactive conversational memory network for multimodal emotion detection. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, 2594–2604.

Hu, D.; Hou, X.; Wei, L.; Jiang, L.; and Mo, Y. 2022. MM-DFN: Multimodal dynamic fusion network for emotion recognition in conversations. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 7037–7041. IEEE.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Jin, Z.; Chen, Y.; Leeb, F.; Gresele, L.; Kamal, O.; Lyu, Z.; Blin, K.; Gonzalez Adauto, F.; Kleiman-Weiner, M.; Sachan, M.; et al. 2023. Cladder: Assessing causal reasoning in language models. *Advances in Neural Information Processing Systems*, 36: 31038–31065.

Kahneman, D. 2011. *Thinking, fast and slow*. macmillan.

Lei, S.; Dong, G.; Wang, X.; Wang, K.; and Wang, S. 2023. InstructERC: Reforming Emotion Recognition in Conversation with a Retrieval Multi-task LLMs Framework. *CoRR*, abs/2309.11911.

Li, Z.; Tang, F.; Zhao, M.; and Zhu, Y. 2022. EmoCaps: Emotion Capsule based Model for Conversational Emotion Recognition. In Muresan, S.; Nakov, P.; and Villavicencio, A., eds., *Findings of the Association for Computational Linguistics: ACL 2022*, 1610–1618. Dublin, Ireland: Association for Computational Linguistics.

Lian, Z.; Liu, B.; and Tao, J. 2021. CTNet: Conversational transformer network for emotion recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29: 985–1000.

Lian, Z.; Sun, H.; Sun, L.; Yi, J.; Liu, B.; and Tao, J. 2024. AffectGPT: Dataset and framework for explainable multimodal emotion recognition. *arXiv preprint arXiv:2407.07653*.

Ma, H.; Wang, J.; Lin, H.; Zhang, B.; Zhang, Y.; and Xu, B. 2023. A transformer-based model with self-distillation for multimodal emotion recognition in conversations. *IEEE Transactions on Multimedia*.

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

Mirza, P.; and Tonelli, S. 2014. An analysis of causality between events and its relation to temporal information. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, 2097–2106.

Nguyen, C.-V. T.; Mai, A.-T.; Le, T.-S.; Kieu, H.-D.; and Le, D.-T. 2023. Conversation understanding using relational temporal graph neural networks with auxiliary cross-modality interaction. *arXiv preprint arXiv:2311.04507*.

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

Rashkin, H.; Sap, M.; Allaway, E.; Smith, N. A.; and Choi, Y. 2018. Event2mind: Commonsense inference on events, intents, and reactions. *arXiv preprint arXiv:1805.06939*.

Sap, M.; Le Bras, R.; Allaway, E.; Bhagavatula, C.; Lourie, N.; Rashkin, H.; Roof, B.; Smith, N. A.; and Choi, Y. 2019. Atomic: An atlas of machine commonsense for if-then reasoning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, 3027–3035.

Satpute, A. B.; Shu, J.; Weber, J.; Roy, M.; and Ochsner, K. N. 2013. The functional neural architecture of self-reports of affective experience. *Biological psychiatry*, 73(7): 631–638.

Shi, T.; and Huang, S.-L. 2023. MultiEMO: An attention-based correlation-aware multimodal fusion framework for emotion recognition in conversations. In *Proceedings of ACL*, 14752–14766.

Shi, T.; Liang, X.; Liang, Y.; Tong, X.; and Huang, S.-L. 2023. SSLCL: An Efficient Model-Agnostic Supervised Contrastive Learning Framework for Emotion Recognition in Conversations. *arXiv preprint arXiv:2310.16676*.

Tu, G.; Liang, B.; Jiang, D.; and Xu, R. 2022. Sentiment-emotion-and context-guided knowledge selection framework for emotion recognition in conversations. *IEEE Transactions on Affective Computing*, 14(3): 1803–1816.

Tu, G.; Liang, B.; Mao, R.; Yang, M.; and Xu, R. 2023. Context or knowledge is not always necessary: A contrastive learning framework for emotion recognition in conversations. In *Findings of the association for computational linguistics: ACL 2023*, 14054–14067.

Tu, G.; Wang, J.; Li, Z.; Chen, S.; Liang, B.; Zeng, X.; Yang, M.; and Xu, R. 2024. Multiple knowledge-enhanced interactive graph network for multimodal conversational emotion recognition. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, 3861–3874.

Tu, G.; Wang, J.; Yang, L.; Liang, B.; Cambria, E.; Li, W.; and Xu, R. 2025. Multi-Task Mutual Learning for Multimodal Emotion-Cause Pair Extraction in Conversations. *Information Fusion*, 103877.

Wang, L.; Wu, J.; Huang, S.-L.; Zheng, L.; Xu, X.; Zhang, L.; and Huang, J. 2019. An efficient approach to informative feature extraction from multimodal data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 5281–5288.

Wu, A.; Kuang, K.; Zhu, M.; Wang, Y.; Zheng, Y.; Han, K.; Li, B.; Chen, G.; Wu, F.; and Zhang, K. 2024. Causality for large language models. *arXiv preprint arXiv:2410.15319*.

Xue, J.; Nguyen, M.-P.; Matheny, B.; and Nguyen, L.-M. 2024. Bioserc: Integrating biography speakers supported by llms for erc tasks. In *International Conference on Artificial Neural Networks*, 277–292. Springer.

Yun, T.; Lim, H.; Lee, J.; and Song, M. 2024. TelME: Teacher-leading Multimodal Fusion Network for Emotion Recognition in Conversation. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 82–95.

Zhang, X.; and Li, Y. 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)*, 13099–13110.

Zheng, X.; Aragam, B.; Ravikumar, P. K.; and Xing, E. P. 2018. Dags with no tears: Continuous optimization for structure learning. *Advances in neural information processing systems*, 31.