UniMEEC: Towards Unified Multimodal Emotion Recognition and Emotion Cause

Guimin Hu[†], **Zhihong Zhu**[♠], **Daniel Hershcovich**[†], **Lijie Hu**[△], **Hasti Seifi**[♡], **Jiayuan Xie**[◇]
[†]University of Copenhagen
[♠]Peking University
[△]King Abdullah University of Science and Technology
[♡]Arizona State University
[◇]The Hong Kong Polytechnic University
rice.hu.x@gmail.com, dh@di.ku.dk, hasti.seifi@asu.edu

Abstract

Multimodal emotion recognition in conversation (MERC) and multimodal emotion-cause pair extraction (MECPE) have recently garnered significant attention. Emotions are the expression of affect or feelings; responses to specific events, or situations – known as emotion causes. Both collectively explain the causality between human emotion and intents. However, existing works treat emotion recognition and emotion cause extraction as two individual problems, ignoring their natural causality. In this paper, we propose a **Uni**fied **M**ultimodal **E**motion recognition and **E**motion-**C**ause analysis framework (**UniMEEC**) to explore the causality between emotion and emotion cause. Concretely, UniMEEC reformulates the MERC and MECPE tasks as mask prediction problems and unifies them with a causal prompt template. To differentiate the modal effects, UniMEEC proposes a multimodal causal prompt to probe the pre-trained knowledge specified to modality and implements cross-task and cross-modality interactions under task-oriented settings. Experiment results on four public benchmark datasets verify the model performance on MERC and MECPE tasks and achieve consistent improvements compared with the previous state-of-the-art methods.

1 Introduction

Recently, multimodal emotion recognition in conversations (MERC) and multimodal emotion-cause pair extraction (MECPE) have attracted increasing attention (Zhang et al., 2021a,b; Hu et al., 2021a,b). Both task play crucial roles in dialog systems, especially in empathetic response generation in a conversation (Fu et al., 2023; Qian et al., 2023; Tian et al., 2022; Hu et al., 2024). MERC detects the emotion category of each utterance in a conversation, while MECPE finds the reasons that trigger a certain emotion for the utterance. Both tasks are tightly related in practice and theory (Baumeister and Cooper, 1981; Dirven, 1997; Russell, 1990; Lee et al., 2019). However, the existing works treat MERC and MECPE as two separate tasks and ignore their causality. On the one hand, emotions are responses to emotion causes (e.g., specific events) (Marks, 1982; Cabanac, 2002). On the other hand, emotion and its emotion causes are interdependent and mutually influential (Russell, 1990; Lee et al., 2019). The two serve as reflections for each other and together provide a causal story of human behavior and intents. Figure 1 illustrates the causal alignment between emotion category and emotion cause (Baumeister and Cooper, 1981; Dirven, 1997).

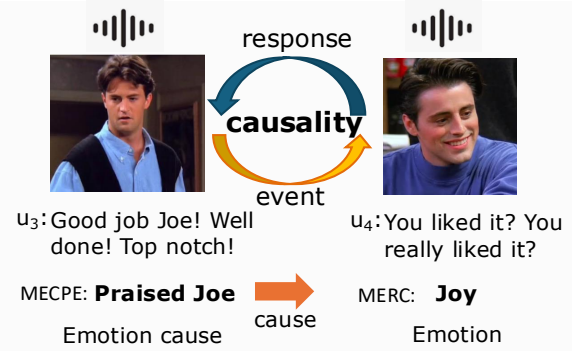

Figure 1: Illustration of the causal inference between emotion and emotion cause, which unifies MECPE and MERC tasks. "response" denotes the speaker's reaction to the event and "event" denotes the event that triggers emotion.

For example, the emotion causes of "happiness" generally are positive events, such as "being praised". Similarly, the emotion causes of "sad" generally are negative events, such as "being criticized". We view the mapping between the specific events (e.g., emotion cause) and response (e.g., emotion label) as the emotion-cause causality. From the causal perspective, Lyu et al. (2024) proposes the idea of causal prompts, which are prompts that describe the causal story behind the

Findings of the Association for Computational Linguistics: EACL 2024, pages 5248–5261

sentiment rating and reviews, further demonstrating that Pretrained Language Model (PLM) is able to be aware of the underlying causality. A natural question arises: *How should we perform causality between emotions and their causes in a unified architecture?*

Recently, the unification of related but different tasks into a framework has achieved significant progress (Chen et al., 2022; Xie et al., 2022; Zhang et al., 2022). For example, UniMSE (Hu et al., 2022b) unifies emotion and sentiment into a single architecture to share complementary knowledge between them. Different from UniMSE which focuses on the unification of emotion and sentiment in a generative way, we propose a multimodal causal prompt to unify MERC and MECPE tasks, thereby capturing the causal nature between emotion and emotion cause. In this paper, we propose a **Uni**fied **M**ultimodal **E**motion recognition and **E**motion-**C**ause pair extraction framework (UniMEEC) to explore the causality between emotion and emotion cause. As Lyu et al. (2024) illustrated, PLM can capture the causal stories with the causal prompts. Starting from this perspective, UniMEEC reformulates MERC and MECPE as two mask prediction tasks and unifies the two tasks using a causal prompt, aiming to capture the understanding of PLM to emotion-cause causlity. In order to differentiate the modal effects, UniMEEC probes modal features from PLM using the multimodal causal prompt, and meanwhile, UniMEEC captures the emotion-specific, cause-specific, and utterance-specific contexts in a hierarchical way. The main contributions are summarized as follows:

- We propose a **Uni**fied **M**ultimodal **E**motion recognition and **E**motion **C**ause pair extraction framework (**UniMEEC**)[1], which uses the causal prompt to unify the MERC and MECPE tasks for causal relation between emotion and emotion cause.
- UniMEEC formalizes MERC and MEEC tasks into mask prediction problems and constructs the multimodal causal prompt to probe the knowledge from PLM. Meanwhile, UniMEEC proposes task-specific context aggregation to orderly capture the contexts oriented to specific tasks.
- Experimental results demonstrate that UniMEEC achieves a new state-of-the-art performance on MELD, IEMOCAP, ConvECPE and ECF datasets, further demonstrating the effectiveness of a unified causal framework for MERC and MECPE.

[1]https://github.com/LeMei/causal-unimeec

2 Related Work

Multimodal Emotion Recognition in Conversations (MERC) We categorize the works of MERC into three main groups: multimodal fusion, context-aware models, and external-knowledge models. The first group focuses on the fusion representation in which some works (Hu et al., 2022a, 2021c; Joshi et al., 2022) employed the graph neural networks to model the inter/intra dependencies of utterances information, and some works proposed cross-attention Transformer (Vaswani et al., 2017) to model cross-modality interaction. Addressing context incorporation, Sun et al. (2021); Li et al. (2021b); Ghosal et al. (2019) construct graph structures to represent contexts and further model inter-utterance dependencies, while Mao et al. (2021) introduces the concept of emotion dynamics to capture context. In the last group, advanced MERC studies integrate external knowledge, employing techniques such as transfer learning (Hazarika et al., 2019; Lee and Lee, 2021), commonsense knowledge (Ghosal et al., 2020), multitask learning (Akhtar et al., 2019), and external information (Zhu et al., 2021) to introduce more auxiliary information to help model understand conversation.

Multimodal Emotion-Cause Pair Extraction (MECPE) As more and more NLP tasks extend to the multimodal paradigm (Zhu et al., 2024; Li et al., 2024; **?**), Wang et al. (2021) defined multimodal emotion-cause pair extraction (MECPE) and constructed Emotion-Cause-in-Friends (ECF) dataset based on MELD (Poria et al., 2019). Li et al. (2022a) built an English conversational emotion-cause pair extraction multimodal dataset based on IEMOCAP (Busso et al., 2008). With MECPE only emerging for a relatively short time, there are a few baseline methods in this field. Previous studie (Wang et al., 2021; Li et al., 2022a) integrated multimodal features to tackle the MECPE task based on the baselines of ECPE (Xia and Ding, 2019), overlooking the importance of inter-utterance context and multimodal fusion in understanding emotion cause.

Prompt-tuning Prompt-tuning (Li and Liang, 2021; Liu et al., 2021; Su et al., 2021), inspired

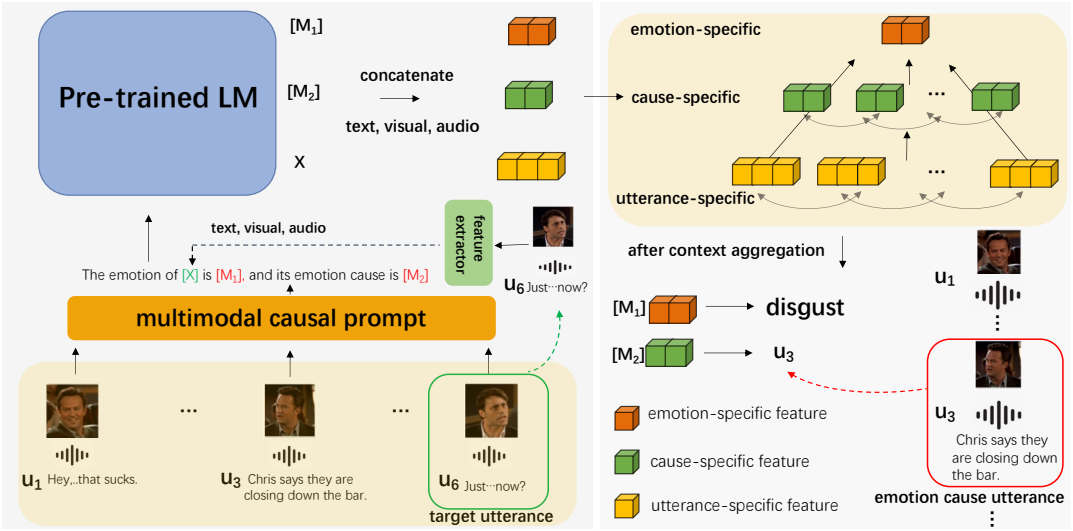

Figure 2: The overview of UniMEEC. The outputs "disgust" and "u_3" denote the emotion category and the emotion cause utterance ID of target utterance u_6, respectively.

by GPT-3 (Ding et al., 2023), is a new paradigm to fine-tuning, particularly geared towards addressing few-shot scenarios. Recently, prompt-tuning has been widely used in addressing NLP tasks and achieved remarkable performances (Zheng et al., 2022; Li et al., 2021a; Yang et al., 2023; Su et al., 2021; Sun et al., 2022). The initial input X undergoes modification through a template to form a textual string prompt X' with unfilled slots. Subsequently, the language model is employed to probabilistically fill in the missing information, resulting in a final string $\hat{X}$ from which the model outputs y (Liu et al., 2023). The prompt template contains manual template engineering and automated template learning (Liu et al., 2023). The manual template is to manually create intuitive templates and the auto-prompt template (Li and Liang, 2021; Liu et al., 2021; Su et al., 2021) includes discrete prompts, represented by actual text strings, and continuous prompts, described directly within the embedding space of the underlying language model. In this work, UniMEEC constructs causal prompts to unify MERC and MECPE, where causal prompt connects emotion and corresponding emotion cause to ensure the causal coherence.

3 Methodology

3.1 Overall Architecture

As shown in Figure 2, UniMEEC is composed of multimodal causal prompt (MCP) and task-specific context aggregation (THC). Multimodal causal prompt template contains modality information [X], auxiliary prompt tokens $\mathrm{P}_{(\cdot)}$, and mask tokens $[\mathrm{M}]_1$ and $[\mathrm{M}]_2$. We feed the causal template into PLM to encode [X], $[\mathrm{M}]_1$ and $[\mathrm{M}]_2$ into vectors. THC takes the emotion-specific, cause-specific, and utterance-specific representations as nodes and models their dependencies in the context window. Finally, UniMEEC predicts the emotion category and the position of cause utterance in a conversation based on the representations of $[\mathrm{M}]_1$ and $[\mathrm{M}]_2$ respectively.

3.2 Task Formalization

Given a multi-turn conversation $U = \{u_1, u_2, \cdots, u_{|U|}\}$, U has $|U|$ utterances and each utterance $u_i = \{I_i^t, I_i^a, I_i^v\}$ contains three modalities, where $I_i^m, m \in \{t, a, v\}$ represent uni-modal feature extracted from video fragment i, and $\{t, a, v\}$ denote the three types of modalities—text, acoustic and visual, respectively. Multimodal emotion recognition (MERC) predicts the emotion category of u_i, and multimodal emotion-cause pair extraction (MECPE) aims to predict the corresponding cause utterance ID (e.g., "u_1", "u_2") for non-neutral utterance u_i. To unify MERC and MECPE, we formalize MERC and MECPE as two mask prediction problems in the causal prompt and leverage the language model to probabilistically fill the unfilled slots, thereby predicting the results of MERC and MECPE tasks respectively.

3.3 Multimodal Causal Prompt (MCP)

In order to differentiate the modal effects, we set causal prompt for each modality to probe the

modality-specific features from PLM. Multimodal causal prompts share auxiliary prompt tokens in the prompt template, which enables inter-modality and inter-task semantic interaction in representation learning.

3.3.1 Causal Prompt Construction

We manually design the modality-specific prompt template, and it consists of a modal input [X], the emotion category slot $[\mathrm{M}]_1$, the cause slot $[\mathrm{M}]_2$ and auxiliary prompt part, where [X] is the slot filled with modal feature of target utterance, $[\mathrm{M}]_1$ indicates the emotion category of target utterance, e.g., "happy" or "sad", and $[\mathrm{M}]_2$ indicates the cause utterance ID of target utterance, e.g., "u_1", "u_2". $[\mathrm{M}]_1$ and $[\mathrm{M}]_2$ are unfilled answer slots and are separately predicted as the results of MERC and MECPE. Given text modality $I_i^t, i \in \{1, \cdots, |U|\}$, we designed the causal prompt template like "the emotion of utterance I_i^t is $[\mathrm{M}]_1$, and its emotion cause is $[\mathrm{M}]_2$" as text-specific prompt, where the textual strings "For conversation", "the emotion category of", "is", and "the reason for this emotion is" are auxiliary prompt parts. For audio-specific and vision-specific prompts, we replace the [X] part of the prompt with the acoustic and visual representations to construct audio-specific and vision-specific prompts, respectively.

We use $X_{i,m}, X_{i,m} \in R^{l_m \times d_m}$ to represent the modal representation after modal alignment (Tsai et al., 2019), l_m and d_m are the sequence length and the representation dimension of modality m, respectively. Specifically, we obtain $X_{i,t}$ with the word embedding layer of the model and we processed raw acoustic input into numerical sequential vectors by librosa [2] to extract Mel-spectrogram as $X_{i,a}$. For vision modality, we use effecientNet (Tan and Le, 2019) pre-trained (supervised) on VGGface [3] and AFEW dataset to extract $X_{i,v}$.

3.3.2 Causal Prompt Encoder

We take Transformer-based model (e.g., BERT (Devlin et al., 2019)) as the backbone of the multimodal causal prompt. The stacked Transformer contains multiple Transformer layers, and each layer contains a self-attention module, FFN, and layer normalization (Ba et al., 2016). We take the former N_t Transformer layers as the text-specific prompt encoder and take the latter N_a and N_v Transformer layers as the visual- and acoustic prompt encoders, respectively. First, text-specific prompt is fed into the text-specific prompt encoder to get the text-specific representations of [X], auxiliary prompt part, and $[\mathrm{M}]_1$ and $[\mathrm{M}]_2$, with the supervision of real ground answers of slots. After that, we obtain the text-specific prompt sequence, which contains the hidden states of $h_{P_{1,l_1}}$, $X_{i,t}$, $h_{P_{l_2,l_3}}$, $h_{[\mathrm{M}]_1}$, $h_{P_{l_4,l_5}}$ and $h_{[\mathrm{M}]_2}$, where $h_{(\cdot)}$ denotes the representation of token or token sequence, $h_{P_{1,l_1}}$, $h_{P_{l_2,l_3}}$ and $h_{P_{l_4,l_5}}$ denote the representations of auxiliary prompt parts.

Due to the dimensions and sequence lengths of audio and vision modalities being less than the dimensions and sequence length of text modality, we pad the audio and vision feature with zero to achieve consistency with the representation of text modality. We take $\hat{X}_{i,a}$ and $\hat{X}_{i,v}$ to represent audio and vision representations after padding, respectively. For audio-specific prompt, we replace [X] part of the prompt representation with $\hat{X}_{i,a}$. For vision-specific prompt, we replace [X] part of the prompt representation with $\hat{X}_{i,v}$ after N_t Transformer layers. After that, we feed audio-specific and vision-specific prompts into N_a and N_v Transformer layers respectively. For (n-1)-th Transformer layer, the modality-specific prompt learning is given by:

$$
\begin{aligned}
P_{i,m}^{n-1} &= [h_{P_{1,l_1}}, X_{i,m}^{n-1}, h_{P_{l_2,l_3}}, h_{[\mathrm{M}]_1}^m, h_{P_{l_4,l_5}}, h_{[\mathrm{M}]_2}^m] \\
P_{i,m}^{n} &= \text{Transformer}(P_{i,m}^{n-1}, P_{i,m}^{n-1}, P_{i,m}^{n-1}) \\
X_{i,m}^{n} &= P_{i,m}^{n}, m \in \{t, a, v\}
\end{aligned} \tag{1}
$$

where $P_{i,m}^{n-1}$ denotes the prompt representation of utterance u_i under the modality m. Specifically, $P_{i,m}^{n-1}$ is composed by the hidden states of [X], $[\mathrm{M}]_1$ $[\mathrm{M}]_2$, and auxiliary prompt strings. $X_{i,t}^0 = X_{i,t}$, $X_{i,a}^0 = \hat{X}_{i,a}$, and $X_{i,v}^0 = \hat{X}_{i,v}$. $[\cdot, \cdot]$ denotes the concatenation operation.

After the multimodal causal prompt, we obtain the modal fusion representations of mask tokens $[\mathrm{M}]_1$ and $[\mathrm{M}]_2$ via concatenation, respectively. Similarly, we obtain the fusion representation of u_i via the concatenation of $X_{i,t}^{N_t}$, $X_{i,a}^{N_a}$ and $X_{i,v}^{N_v}$:

$$
\begin{aligned}
h_{[\mathrm{M}]_1}^f &= [h_{[\mathrm{M}]_1}^t, h_{[\mathrm{M}]_1}^a, h_{[\mathrm{M}]_1}^v] \\
h_{[\mathrm{M}]_2}^f &= [h_{[\mathrm{M}]_2}^t, h_{[\mathrm{M}]_2}^a, h_{[\mathrm{M}]_2}^v] \\
h_{u_i}^f &= [X_{i,t}^{N_t}, X_{i,a}^{N_a}, X_{i,v}^{N_v}]
\end{aligned} \tag{2}
$$

where $X_{i,t}^{N_t}$, $X_{i,a}^{N_a}$ and $X_{i,v}^{N_v}$ are text, audio and video representations of u_i encoded by N_t, N_a and N_v Transformer layers respectively.

[2] https://github.com/librosa/librosa

[3] https://www.robots.ox.ac.uk/~vgg/software/vgg_face/

3.4 Task-specific Hierarchical Context (THC)

The learned representations of $[\mathrm{M}]_1$ (i.e., $h^f_{[\mathrm{M}]_1}$) and $[\mathrm{M}]_2$ (i.e., $h^f_{[\mathrm{M}]_2}$) fail to capture the context information in a conversation, which inspires us to build a hierarchical context aggregation structure to control the direction of context aggregation in a conversation. In order to avoid the noise information in representation learning, we set the context windows for each utterance to incorporate the information around target utterance.

3.4.1 Hierarchical Graph Construction

We construct a 3-level graph attention network (GAT) (Velickovic et al., 2018) as the encoder of contexts, which includes top, middle, and bottom levels. Each level has a context window to focus on the local context of utterance. Formally, we define a graph $G = (V, E)$, V and E denote the node and edge sets respectively. We take the utterance-level representation h_u as the bottom node, cause-specific token representation $h^f_{[\mathrm{M}]_2}$ as the middle node, and the emotion-specific token representation $h^f_{[\mathrm{M}]_1}$ as the top node. For the intra-level nodes, we set undirected edges for any two adjacent nodes in the context window of the same level. For the inter-level nodes, we set the undirected edges between the top nodes and middle nodes. In general, we set the directed edges from the bottom to the middle nodes in the context window, aiming to control the direction of the information flow among nodes.

Considering that graph G contains multiple type node representations, we set five edge types respectively to model the dependency relations among different nodes. The former three edges are constructed between the slot nodes to slot nodes, i.e., $h_{[\mathrm{M}]_1} \leftrightarrow h_{[\mathrm{M}]_1}$, $h_{[\mathrm{M}]_1} \leftrightarrow h_{[\mathrm{M}]_2}$ and $h_{[\mathrm{M}]_2} \leftrightarrow h_{[\mathrm{M}]_2}$, which are represented with t_{ee}, t_{ec} and t_{cc} respectively. The fourth edge type is constructed from utterance node to slot node, i.e., $h_u \leftrightarrow h_{[\mathrm{M}]_2}$, represented by t_{uc}. The last is from utterance node to utterance node, i.e., $h_u \leftrightarrow h_u$, denoted by t_{uu}. The subscripts "e" and "c" in edge type represent $[\mathrm{M}]_1$ and $[\mathrm{M}]_2$, respectively, and "u" represents the utterance. For one edge type $t \in \{t_{ee}, t_{ec}, t_{cc}, t_{uc}, t_{uu}\}$, its adjacent matrix is given as:

$$a^t_{i,j} = \begin{cases} 1 & j \in \{i - |w|, i + |w|\} \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

where $a^t_{i,j} \in A, A \in R^{V*V}$. V denotes the number of utterances in a conversation. $|w|$ denotes the size of the context window. i and j represent the indexes of utterances in a conversation, and they are located on the same or adjacent levels of THC.

3.4.2 Task-specific Context Aggregation

We set a contextual window for each node at each level to ensure that the model only aggregates the node representations in its contextual window. This operation reduces the computational cost and avoids introducing noise to the representation learning. Given an utterance u_i, the prediction slots of emotion and emotion cause are $[\mathrm{M}]_{i,1}$ and $[\mathrm{M}]_{i,2}$ respectively. We aggregate the representation from the bottom to top levels in the graph, and the representations of bottom nodes are not updated by aggregating the representations of the top or middle nodes to them. For the bottom node u_i, its representation is aggregated by the bottom nodes in the context window:

$$h^n_{u_i} = \mathrm{ReLU}\left(\sum_{j \in \mathcal{N}_{u_i}} a^{t_{uu}}_{i,j} W^{uu,n-1} h^{n-1}_{u_j} + b^{n-1}\right) \tag{4}$$

where $\mathcal{N}_{u_i}$ denotes the neighbor nodes of utterance u_i and $h^0_{u_j} = h^f_{u_j}$. When the model comes to the middle node $[\mathrm{M}]_{i,2}$, the representations is aggregated by the top and middle nodes in the context window, which is given by:

$$\begin{aligned} h^n_{[\mathrm{M}]_{i,2}} = \mathrm{ReLU}(& \sum_{j \in \mathcal{N}_{[\mathrm{M}]_{i,2}}} a^{t_{cc}}_{i,j} W^{cc,n-1} h^{n-1}_{[\mathrm{M}]_{j,2}} \\ & + \sum_{j \in \mathcal{N}_{[\mathrm{M}]_{i,1}}} a^{t_{ec}}_{i,j} W^{m_{ec},n-1} h^{n-1}_{[\mathrm{M}]_{j,1}}) \\ & + \sum_{j \in \mathcal{N}_{u_i}} a^{t_{uc}}_{i,j} W^{uc,n-1} h^{n-1}_{u_j} + b^{n-1}) \end{aligned} \tag{5}$$

where $\{\mathcal{N}_{[\mathrm{M}]_{i,1}}, \mathcal{N}_{[\mathrm{M}]_{i,2}}\}$ denote the neighbor nodes of tokens $[\mathrm{M}]_1$ and $[\mathrm{M}]_2$ respectively. $h^0_{[\mathrm{M}]_{j,1}} = h^f_{[\mathrm{M}]_{j,1}}$, $h^0_{[\mathrm{M}]_{j,2}} = h^f_{[\mathrm{M}]_{j,2}}$. When the model comes to the top node $[\mathrm{M}]_{i,1}$, its representation is aggregated by the top, and the middle nodes in the context window, which is given by:

$$\begin{aligned} h^n_{[\mathrm{M}]_{i,1}} = \mathrm{ReLU}(& \sum_{j \in \mathcal{N}_{[\mathrm{M}]_{i,1}}} a^{t_{ee}}_{i,j} W^{ee,n-1} h^{n-1}_{[\mathrm{M}]_{j,1}} \\ & + \sum_{j \in \mathcal{N}_{[\mathrm{M}]_{i,2}}} a^{t_{ec}}_{i,j} W^{ec,n-1} h^{n-1}_{[\mathrm{M}]_{j,2}} + b^{n-1}) \end{aligned} \tag{6}$$

We stacked N task-specific context aggregation modules and then use $h^N_{[\mathrm{M}]_{i,1}}$ and $h^N_{[\mathrm{M}]_{i,2}}$ as final representations of slots $[\mathrm{M}]_{i,1}$ and $[\mathrm{M}]_{i,2}$ respectively.

Datasets	Train	Valid	Test	All
MELD	9989	1108	2610	13707
IEMOCAP	5354	528	1650	7532
ConvECPE	5303	486	1644	7433
ECF	9457	1351	2701	13509

Table 1: The statistics of MELD, IEMOCAP, ConvECPE, and ECF.

3.5 Grounding Mask Predictions to MERC and MECPE

We use $h^N_{[\mathrm{M}]_{i,1}}$ to predict MERC task, i.e., the answers of slot $[\mathrm{M}]_1$, and use $h^N_{[\mathrm{M}]_{i,2}}$ to predict MECPE task, i.e., the answers of slot $[\mathrm{M}]_2$. The predictions of $[\mathrm{M}]_1$ (i.e., $\hat{y}^e_i$) and $[\mathrm{M}]_1$ (i.e., $\hat{y}^c_i$) are given as respectively:

$$\begin{aligned} \hat{y}^e_i &= f(W^e h^N_{[\mathrm{M}]_{i,1}} + b^e) \\ \hat{y}^c_i &= f(W^c h^N_{[\mathrm{M}]_{i,2}} + b^c) \end{aligned} \tag{7}$$

where $\{\hat{y}^e_i, \hat{y}^c_i\}$ denote the prediction results for MERC and MECPE tasks, respectively. Based on the predictions, we use the sum of the cross-entropy losses of MERC and MECPE tasks as the objective loss of UniMEEC.

4 Experiments

4.1 Datasets

We conduct experiments on four publicly available benchmark datasets of MERC and MECPE. For MERC task, its benchmark datasets include multimodal emotionLines dataset (**MELD**) (Poria et al., 2019), interactive emotional dyadic motion capture database (**IEMOCAP**) (Busso et al., 2008). **IEMOCAP** consists of 7532 samples, and each sample is labeled with six emotions for emotion recognition, including happiness, sadness, anger, neutral, excitement, and frustration. **MELD** contains 13,707 video clips of multi-party conversations, with labels following Ekman's six universal emotions, including joy, sadness, fear, angry, surprise and disgust. For MECPE task, its benchmark datasets include **ConvECPE** (Li et al., 2022a), and emotion-cause-in-friends (**ECF**) (Wang et al., 2021). **ConvECPE** is a multimodal emotion cause dataset constructed based on IEMOCAP, in which each non-neutral utterance is labeled with the emotion cause. It contains 151 dialogues with 7,433 utterances. Similarly, (Wang et al., 2021) annotated the emotion cause of each sample in MELD and then constructed multimodal emotion cause dataset ECF. **ECF** contains 1,344 conversations and 13,509 utterances. The detailed statistics of four datasets are shown in Table 1. For datasets IEMOCAP and MELD, we follow previous works (Li et al., 2021c; Lu et al., 2020), and we use accuracy (ACC) and weighted F1 (WF1) as the evaluation metric for the MERC task. For datasets ECF and ConvECPE, we use precision (P), recall (R), and F1 as the evaluation metric for the MECPE task.

4.2 Baselines

For MERC, the baselines can be grouped into three categories: 1)the methods focusing on emotion cues like **EmoCaps** (Li et al., 2022b), **FacialMMT-RoBERTa** (Zheng et al., 2023), **MVN** (Li et al., 2021c). These works aim to improve model performance by tracking emotional states in a conversation, and 2)the methods fusing multimodal information like **QMNN** (Li et al., 2021c), **GA2MIF** (Li et al., 2023),**MALN**(Ren et al., 2023), **MultiEMO** (Shi and Huang, 2023), and **UniMSE** (Hu et al., 2022b). These works focus on better multimodal fusion, and 3)the methods incorporating context information like **DialogueGCN** (Ghosal et al., 2019), **MMGCN** (Hu et al., 2021c), **MM-DFN** (Hu et al., 2022a), **BC-LSTM** (Poria et al., 2017), **DialogueRNN** (Majumder et al., 2019) and **IterativeERC** (Lu et al., 2020). These works aggregate the context to understand the whole conversation.

MECPE has a few baselines due to MECPE only emerging for a relatively short time. Most baselines address MECPE tasks based on two-step frameworks of emotion-cause pair extraction in text, like **Joint-GCN** (Li et al., 2022a), **Joint-Xatt**(Li et al., 2022a) and **Inter-EC**(Li et al., 2022a). $\mathbf{C}_{\textbf{Multi-Bernoulli}}$(Wang et al., 2021) carries out a binary decision for each relative position to determine the cause utterance. $\mathbf{C}_{\textbf{Multinomial}}$ (Wang et al., 2021) randomly selects a relative position from all relative positions as the feature to extract emotion-cause pair. We produce some typical multimodal methods based on their open source codes, including **MuLT** (Tsai et al., 2019), **MMGCN** (Hu et al., 2021c), **MMDFN** (Hu et al., 2022a), **UniMSE** (Hu et al., 2022b) and **GA2MIF** (Li et al., 2023).

4.3 Experimental Settings

We use pre-trained BERT as the encoder of multimodal causal prompt. ConvECPE and ECF are constructed based on IEMOCAP and MELD re-

Methods	IEMOCAP							MELD							
	Happiness	**Sadness**	**Neutral**	**Anger**	**Excitement**	**Frustration**	**WF1**	**Neutral**	**Surprise**	**Fear**	**Sadness**	**Joy**	**Disgust**	**Angry**	**WF1**
BC-LSTM(Poria et al., 2017)	34.43	60.87	51.81	56.73	57.95	58.92	54.95	73.80	47.70	5.40	25.1	51.30	5.20	38.40	55.90
DialogueRNN(Majumder et al., 2019)	33.18	78.80	59.21	65.28	71.86	58.91	62.75	76.23	49.59	0.00	26.33	54.55	0.81	46.76	58.73
DialogueGCN(Ghosal et al., 2019)	51.87	76.76	56.76	62.26	72.71	58.04	63.16	76.02	46.37	0.98	24.32	53.62	1.22	43.03	57.52
IterativeERC(Lu et al., 2020)	53.17	77.19	61.31	61.45	69.23	60.92	64.37	77.52	53.65	3.31	23.62	56.63	19.38	48.88	60.72
QMNN(Li et al., 2021c)	39.71	68.30	55.29	62.58	66.71	62.19	59.88	77.00	49.76	0.00	16.50	52.08	0.00	43.17	58.00
MMGCN(Hu et al., 2021c)	42.34	78.67	61.73	69.00	74.33	62.32	66.22	-	-	-	-	-	-	-	58.65
MM-DFN(Hu et al., 2022a)	42.22	78.98	66.42	69.77	75.56	66.33	68.18	77.76	50.69	-	22.93	54.78	-	47.82	58.65
MVN(Ma et al., 2022)	55.75	73.30	61.88	65.96	69.50	64.21	65.44	76.65	53.18	11.70	21.82	53.62	21.86	42.55	59.03
UniMSE(Hu et al., 2022b)	-	-	-	-	-	-	70.66	-	-	-	-	-	-	-	65.51
EmoCaps(Li et al., 2022b)	**71.91**	85.06	64.48	68.99	78.41	66.76	71.77	77.12	63.19	3.03	42.52	57.50	7.69	57.54	64.00
GA2MIF(Zheng et al., 2023)	46.15	84.50	68.38	70.29	75.99	66.49	70.00	76.92	49.08	-	27.18	51.87	-	48.52	58.94
FacialMMT-RoBERTa(Zheng et al., 2023)	-	-	-	-	-	-	-	80.13	59.63	19.18	41.99	64.88	18.18	56.00	66.58
MALN(Ren et al., 2023)	55.50	81.80	64.10	69.10	78.00	71.40	70.80	82.00	58.60	21.20	43.00	64.30	17.60	52.40	66.90
MultiEMO(Shi and Huang, 2023)	65.77	85.49	67.08	69.88	77.31	70.98	72.84	79.95	60.98	29.67	41.51	62.82	36.75	54.41	66.74
UniMEEC (Ours)	69.52	**88.51**	**69.74**	**72.63**	**78.80**	**72.98**	**74.83**	**82.75**	**64.28**	**31.78**	**43.31**	**66.91**	**37.72**	**58.46**	**68.96**

Table 2: Results on IEMOCAP and MELD datasets. The best results are highlighted in bold. The results with underline denote the previous SOTA performance.

	IEMOCAP		MELD	
	ACC	WF1	ACC	WF1
BART	73.59	74.46	74.69	68.84
T5	74.32	75.09	74.93	69.06
LLaMA	74.67	75.16	75.02	69.15

Table 3: Experimental results on IEMOCAP and MELD datasets with BART, T5 and LLaMA as backbone.

spectively, so we integrate the emotion and cause labels of IEMOCAP, MELD, ConvECPE and ECF to train the model. The batch size is 64, the learning rate for BERT fine-tuning is set at 3e-4, and the learning rate for UniMEEC is set to 0.0001. The hidden dimension of acoustic and visual representation is 64, the BERT embedding size is 768, and the fusion vector size is 768. We use the former 9 Transformer layers of BERT as the text-specific prompt encoder, the following 10th and 11th as the audio-specific prompt encoder, and the last Transformer layer of BERT as the video-specific prompt encoder. The THC module stacks two graph network layers, where the first layer has one attention head and the second layer has four attention heads.

4.4 Experimental Environment

All experiments are conducted in the NVIDIA RTX A100. We take BERT as the Transformer-based model, which has 110M parameters, including 12 layers, 768 hidden dimensions, and 12 heads. We use the former $N_t = 9$ Transformer layers as the text-specific encoder, use the following $N_a = 2$ and $N_v = 1$ Transformer layers as the audio-specific and video-specific encoders respectively. The value of N_t, N_a and N_v are determined by the model performance on valid test. Furthermore, we employ a linear decay learning rate schedule with a warm-up strategy.

4.5 Results of Emotion Recognition

We compare UniMEEC with the baselines of MERC on IEMOCAP and MELD datasets, and the comparative results are shown in Table 2. UniMEEC significantly outperforms SOTA in all metrics on IEMOCAP, and MELD, and improves WF1 scores of IEMOCAP and MELD by 1.99% and 2.06%, respectively.

Recent methods like MultiEMO, MALN, and GA2MF achieve low performance in recognizing the label "Happiness" for the IEMOCAP dataset and recognizing the label "Fear" for the MELD dataset. The low performance is caused by the label imbalance of the benchmark. UniMEEC significantly improves the emotion recognition performance on most emotion categories for two datasets. On the one hand, the unified framework offers model auxiliary information, enhancing the interaction between emotion and emotion cause, thereby alleviating the label imbalance of the benchmark. On the other hand, UniMEEC unifies the annotated labels of MERC and MECPE tasks with a causal prompt, which probes the causal story between response (emotion) and event (emotion cause). In summary, UniMEEC consistently surpasses the state-of-the-art (SOTA) in most emotion category recognition on both datasets. These results indicate the superiority of UniMEEC to MERC and MECPE and illustrate the unified framework of modeling emotion-cause causality brings improvements to emotion recognition.

Furthermore, we explore the impact of different PLMs, i.e., BART (Lewis et al., 2020), T5 (Raffel et al., 2020) and LLaMa (Touvron et al., 2023) on UniMEEC performance. We report the result on IEMOCAP and MELD datasets when we take BART, T5 and LLaMA as the PLM of UniMEEC. The experimental results are shown in Table 3.

Methods	Cause Recognition			Pair Extraction			
	P	R	F1	P	R	F1	WF1
E_{True} + $C_{Multi\text{-}Bernoulli}$(Wang et al., 2021)	55.69	57.20	55.47	49.40	25.22	33.39	-
E_{True} +$C_{Multinomial}$(Wang et al., 2021)	57.21	56.38	56.85	49.33	25.18	33.34	-
MC-ECPE-2steps(Wang et al., 2021)	57.76	56.71	57.09	49.43	53.76	51.32	30.00
MuLT*(Tsai et al., 2019)	55.19	53.43	54.79	30.48	37.85	39.02	-
MMGCN*(Hu et al., 2021c)	56.51	54.82	55.30	35.43	38.19	37.48	54.65
MM-DFN*(Hu et al., 2022a)	54.28	56.35	55.17	37.90	39.08	38.10	54.86
UniMSE*(Hu et al., 2022b)	56.55	57.09	56.73	44.48	54.25	49.08	56.37
GA2MIF*(Zheng et al., 2023)	56.48	58.33	56.67	46.15	54.26	50.16	57.33
UniMEEC(Ours)	**59.87**	**58.85**	**59.18**	**49.88**	**59.29**	**54.61**	**63.67**

Table 4: Results on ECF dataset. Cause recognition is to predict the location of cause utterance and pair extraction is to match the emotion utterance and cause utterance, and WF1 denotes the performance of emotion recognition. The baselines with * are reproduced with their open sources.

Methods	Cause Recognition			Pair Extraction			
	P	R	F1	P	R	F1	WF1
Joint-GCN(Joint-EC)(Li et al., 2022a)	71.47	86.35	78.21	38.23	37.08	37.65	-
Joint-Xatt(Joint-EC)(Li et al., 2022a)	69.68	89.42	78.33	38.23	37.08	37.65	-
Inter-EC(Li et al., 2022a)	68.55	85.55	76.11	30.91	37.34	33.82	-
MuLT*(Tsai et al., 2019)	75.15	71.43	73.05	44.61	**52.59**	48.74	-
MMGCN*(Hu et al., 2021c)	78.57	74.52	76.07	42.18	42.67	42.11	63.28
MM-DFN*(Hu et al., 2022a)	79.84	74.11	76.90	46.79	50.36	48.50	65.51
UniMSE*(Hu et al., 2022b)	80.37	73.09	75.58	44.24	49.33	46.69	67.36
GA2MIF*(Zheng et al., 2023)	81.42	75.36	78.71	46.54	48.59	47.40	-
UniMEEC(Ours)	**87.21**	**92.95**	**89.88**	**50.61**	50.41	**50.83**	**69.48**

Table 5: Results on ConvECPE dataset. The baselines with italics indicate it only uses textual modality.

4.6 Results of Emotion-Cause Pair Extraction

The results of cause recognition, pair extraction, and emotion recognition on ECF and ConvECPE datasets are shown in Table 4 and Table 5, respectively. UniMEEC significantly outperforms SOTA in all metrics on ECF and most metrics on ConvECPE datasets. For the ECF dataset, UniMEEC improves metrics P, R, and F of cause recognition by 2.11%, 0.52%, and 2.09%, respectively, and P, R, and F of pair recognition by 0.45%, 5.03%, and 3.29% respectively. For the ConvECPE dataset, multimodal methods perform better than text-based ones. UniMEEC improves by at least 2% on most metrics for cause recognition and pair extraction. Furthermore, we report the UniMEEC performance of the emotion recognition task on two datasets (see WF1 in Table 4 and Table 5), outperforming at least 5.34% and 2.12% improvements by the competitive baselines on ECF and ConvECPE, respectively.

We summarize the improvements into two aspects: 1) UniMEEC achieves SOTA on emotion recognition, cause recognition, and emotion-cause pair extraction on the benchmarks of MERC and MECPE, and 2)UniMEEC significantly outperforms SOTA in most cases. The improvements illustrate jointly training emotion and emotion cause can benefit the two tasks, and the unified framework in modeling causality between emotion and emotion cause can bring prior knowledge to MERC and MECPE training.

4.7 Ablation Study

We conducted extensive ablation studies on IEMOCAP and MELD datasets and experimental results are shown in Table 6. First, we remove the MECPE part in the prompt template, and then train UniMEEC just using the emotion label as the supervision signal. The removal of MECPE from UniMEEC results in a performance drop by 3.57% and 1.96% on IEMOCAP and MELD respectively, demonstrating that jointly training MERC and MECPE can bring improvements for

MERC tasks.

Then we remove one or two modalities from MCP by replacing MCP with unimodal and bimodal prompt templates, where unimodal and bimodal prompt templates denote the prompt template containing one and two modalities, respectively. We feed the unimodal and bimodal prompts into PLM and their performances significantly decline on two datasets. We can find that removing acoustic, visual, and textual modalities or one of them all leads to performance degradation, further demonstrating the effectiveness and necessity of multimodal prompt learning to model performance. For example, we eliminate acoustic, visual, and both modalities from the multimodal prompt template, resulting in performance degradation by 2.75%, 1.96%, and 3.56%, respectively, on WF1 for IEMOCAP. Similarly, the performance also drops for the MELD dataset after removing acoustic, visual, and both. For the context aggregation module, we first remove THC from the model, which leads to 1.99% and 3.54% drops on two datasets respectively. Next, we disorder the positions of utterance-specific, cause-specific, and emotion-specific nodes in the THC module, disrupting the hierarchical structure of context aggregation, which results in 1.79% and 1.94% drops on IEMOCAP and MELD respectively. Additionally, It can be found that removing the restriction of the context window when we construct the edges between nodes leads to the drop in ACC and WF1 on two datasets. Overall, MCP and THC are necessary to improve model performance, and introducing MERC and MECPE into a unified framework can bring improvements.

		IEMOCAP		MELD	
		ACC	WF1	ACC	WF1
Task	- w/o MECPE	68.55	71.26	71.41	66.79
UPL	- w/o MCP	68.04	72.70	71.52	65.32
	- w/o A, T	68.02	71.19	69.74	62.96
	- w/o A, V	69.37	72.84	70.65	65.05
	- w/o T, V	68.59	71.88	69.86	63.24
BPL	- w/o A	70.19	72.08	73.42	65.66
	- w/o V	71.02	72.87	73.65	66.89
	- w/o T	67.75	71.23	69.76	65.47
Context	- w/o THC	69.16	72.84	71.09	65.28
	- w/o hierarchy	69.97	73.04	71.76	66.81
	- w/o $\|w\|$	57.38	59.14	58.62	56.30
UniMEEC (Ours)		**73.67**	**74.83**	**74.85**	**68.75**

Table 6: Ablation study of UniMEEC on IEMOCAP and MELD datasets. T, V and A represent textual, visual and acoustic modalities, respectively. UPL and BPL denotes unimodal and bimodal causal prompts, respectively. Hierarchy denotes the hierarchical structure of THC.

5 Conclusion

This paper presents a unified multimodal emotion recognition and emotion-cause analysis framework, which aims to explore the emotion-cause causality by jointly modeling multimodal emotion recognition and emotion-cause pair extraction. UniMEEC reformulates MERC and MECPE tasks as two mask prediction problems, tunes PLM via multimodal causal prompts specific to uni-modality, and aggregates task-specific context in a conversation. Experiments on IEMOCAP, MELD, ConvECPE, and ECF consistently gain significant improvements on most metrics compared to the previous SOTA, further demonstrating the effectiveness of UniMEEC in addressing MERC and MEPCE.

Limitations

Due to the dimensions and sequence lengths of audio and vision modalities being less than the dimensions and sequence length of text modality, UniMEEC pads the audio and vision feature with zero to achieve consistency with the representation of text modality. This operation might introduce some unnecessary information in fusion representation learning. Furthermore, UniMEEC is set up to detect emotion and emotion cause in multimodal scenarios, fails to effectively address MERC and MECPE in text, which will also be solved in our future work.

Ethics Statement

The data used in this study are all open-source data for research purposes. While making machines understand human emotions and behaviors sounds appealing, it could be applied to emotional companion robots or intelligent customer service. However, even in simple multi-class emotion recognition , the proposed method can achieve only 74% and 68% in accuracy on IEMOCAP and MELD respectively, which is far from usable in real-world application.

Acknowledgement

This work was supported by research grants from VILLUM FONDEN (VIL50296) and the National Science Foundation (#2339707).

References

Md. Shad Akhtar, Dushyant Singh Chauhan, Deepanway Ghosal, Soujanya Poria, Asif Ekbal, and Push-

pak Bhattacharyya. 2019. Multi-task learning for multi-modal emotion recognition and sentiment analysis. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 370–379. Association for Computational Linguistics.

Lei Jimmy Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. 2016. Layer normalization. *CoRR*, abs/1607.06450.

Roy F Baumeister and Joel Cooper. 1981. Can the public expectation of emotion cause that emotion? 1. *Journal of Personality*, 49(1):49–59.

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. *Lang. Resour. Evaluation*, 42(4):335–359.

Michel Cabanac. 2002. What is emotion? *Behavioural processes*, 60(2):69–83.

Zhi Chen, Lu Chen, Bei Chen, Libo Qin, Yuncong Liu, Su Zhu, Jian-Guang Lou, and Kai Yu. 2022. Unidu: Towards A unified generative dialogue understanding framework. *CoRR*, abs/2204.04637.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Bosheng Ding, Chengwei Qin, Linlin Liu, Yew Ken Chia, Boyang Li, Shafiq Joty, and Lidong Bing. 2023. Is GPT-3 a good data annotator? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 11173–11195.

René Dirven. 1997. Emotions as cause and the cause of emotions. *The language of emotions*, pages 55–83.

Fengyi Fu, Lei Zhang, Quan Wang, and Zhendong Mao. 2023. E-core: Emotion correlation enhanced empathetic dialogue generation. *arXiv preprint arXiv:2311.15016*.

Deepanway Ghosal, Navonil Majumder, Alexander F. Gelbukh, Rada Mihalcea, and Soujanya Poria. 2020. COSMIC: commonsense knowledge for emotion identification in conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2020, Online Event, 16-20 November 2020*, pages 2470–2481.

Deepanway Ghosal, Navonil Majumder, Soujanya Poria, Niyati Chhaya, and Alexander F. Gelbukh. 2019. Dialoguegcn: A graph convolutional neural network for emotion recognition in conversation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 154–164. Association for Computational Linguistics.

Devamanyu Hazarika, Soujanya Poria, Roger Zimmermann, and Rada Mihalcea. 2019. Emotion recognition in conversations with transfer learning from generative conversation modeling. *CoRR*, abs/1910.04980.

Dou Hu, Xiaolong Hou, Lingwei Wei, Lian-Xin Jiang, and Yang Mo. 2022a. MM-DFN: multimodal dynamic fusion network for emotion recognition in conversations. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2022, Virtual and Singapore, 23-27 May 2022*, pages 7037–7041.

Guimin Hu, Ting-En Lin, Yi Zhao, Guangming Lu, Yuchuan Wu, and Yongbin Li. 2022b. Unimse: Towards unified multimodal sentiment analysis and emotion recognition. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 7837–7851.

Guimin Hu, Guangming Lu, and Yi Zhao. 2021a. Bidirectional hierarchical attention networks based on document-level context for emotion cause extraction. In *Findings of the Association for Computational Linguistics: EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 16-20 November, 2021*, pages 558–568.

Guimin Hu, Guangming Lu, and Yi Zhao. 2021b. FSS-GCN: A graph convolutional networks with fusion of semantic and structure for emotion cause analysis. *Knowl. Based Syst.*, 212:106584.

Guimin Hu, Yi Xin, Weimin Lyu, Haojian Huang, Chang Sun, Zhihong Zhu, Lin Gui, and Ruichu Cai. 2024. Recent trends of multimodal affective computing: A survey from nlp perspective. *arXiv preprint arXiv:2409.07388*.

Jingwen Hu, Yuchen Liu, Jinming Zhao, and Qin Jin. 2021c. MMGCN: multimodal fusion via deep graph convolution network for emotion recognition in conversation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 5666–5675. Association for Computational Linguistics.

Abhinav Joshi, Ashwani Bhat, Ayush Jain, Atin Vikram Singh, and Ashutosh Modi. 2022. COGMEN: contextualized GNN based multimodal emotion recognition. *CoRR*, abs/2205.02455.

Joosung Lee and Wooin Lee. 2021. Compm: Context modeling with speaker's pre-trained memory tracking for emotion recognition in conversation. *CoRR*, abs/2108.11626.

S Lee, Sophia Yat Mei Lee, and Zhu. 2019. *Emotion and Cause*. Springer.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 7871–7880.

Chengxi Li, Feiyu Gao, Jiajun Bu, Lu Xu, Xiang Chen, Yu Gu, Zirui Shao, Qi Zheng, Ningyu Zhang, Yongpan Wang, and Zhi Yu. 2021a. Sentiprompt: Sentiment knowledge enhanced prompt-tuning for aspect-based sentiment analysis. *CoRR*, abs/2109.08306.

Jiang Li, Xiaoping Wang, Guoqing Lv, and Zhigang Zeng. 2023. Ga2mif: Graph and attention based two-stage multi-source information fusion for conversational emotion detection. *IEEE Transactions on Affective Computing*.

Jiangnan Li, Zheng Lin, Peng Fu, and Weiping Wang. 2021b. Past, present, and future: Conversational emotion recognition through structural modeling of psychological knowledge. In *Findings of the Association for Computational Linguistics: EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 16-20 November, 2021*, pages 1204–1214. Association for Computational Linguistics.

Qiuchi Li, Dimitris Gkoumas, Alessandro Sordoni, Jian-Yun Nie, and Massimo Melucci. 2021c. Quantum-inspired neural network for conversational emotion recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 13270–13278.

Wei Li, Yang Li, Vlad Pandelea, Mengshi Ge, Luyao Zhu, and Erik Cambria. 2022a. Ecpec: Emotion-cause pair extraction in conversations. *IEEE Transactions on Affective Computing*, pages 1–12.

Wenyan Li, Xinyu Zhang, Jiaang Li, Qiwei Peng, Raphael Tang, Li Zhou, Weijia Zhang, Guimin Hu, Yifei Yuan, Anders Søgaard, et al. 2024. Foodieqa: A multimodal dataset for fine-grained understanding of chinese food culture. *arXiv preprint arXiv:2406.11030*.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 4582–4597.

Zaijing Li, Fengxiao Tang, Ming Zhao, and Yusen Zhu. 2022b. Emocaps: Emotion capsule based model for conversational emotion recognition. *arXiv preprint arXiv:2203.13504*.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35.

Xiao Liu, Kaixuan Ji, Yicheng Fu, Zhengxiao Du, Zhilin Yang, and Jie Tang. 2021. P-tuning v2: Prompt tuning can be comparable to fine-tuning universally across scales and tasks. *CoRR*, abs/2110.07602.

Xin Lu, Yanyan Zhao, Yang Wu, Yijian Tian, Huipeng Chen, and Bing Qin. 2020. An iterative emotion interaction network for emotion recognition in conversations. In *Proceedings of the 28th international conference on computational linguistics*, pages 4078–4088.

Zhiheng Lyu, Zhijing Jin, Fernando Gonzalez, Rada Mihalcea, Bernhard Schölkopf, and Mrinmaya Sachan. 2024. *CoRR*, abs/2404.11055.

Hui Ma, Jian Wang, Hongfei Lin, Xuejun Pan, Yijia Zhang, and Zhihao Yang. 2022. A multi-view network for real-time emotion recognition in conversations. *Knowledge-Based Systems*, 236:107751.

Navonil Majumder, Soujanya Poria, Devamanyu Hazarika, Rada Mihalcea, Alexander F. Gelbukh, and Erik Cambria. 2019. Dialoguernn: An attentive RNN for emotion detection in conversations. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 6818–6825.

Yuzhao Mao, Guang Liu, Xiaojie Wang, Weiguo Gao, and Xuan Li. 2021. Dialoguetrm: Exploring multi-modal emotional dynamics in a conversation. In *Findings of the Association for Computational Linguistics: EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 16-20 November, 2021*, pages 2694–2704.

Joel Marks. 1982. A theory of emotion. *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition*, 42(2):227–242.

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, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 873–883.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2019. MELD: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 527–536. Association for Computational Linguistics.

Yushan Qian, Bo Wang, Ting-En Lin, Yinhe Zheng, Ying Zhu, Dongming Zhao, Yuexian Hou, Yuchuan Wu, and Yongbin Li. 2023. Empathetic response generation via emotion cause transition graph. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67.

Minjie Ren, Xiangdong Huang, Jing Liu, Ming Liu, Xuanya Li, and An-An Liu. 2023. MALN: multimodal adversarial learning network for conversational emotion recognition. *IEEE Trans. Circuits Syst. Video Technol.*, 33(11):6965–6980.

James A Russell. 1990. The preschooler's understanding of the causes and consequences of emotion. *Child Development*, 61(6):1872–1881.

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), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 14752–14766.

Yusheng Su, Xiaozhi Wang, Yujia Qin, Chi-Min Chan, Yankai Lin, Huadong Wang, Kaiyue Wen, Zhiyuan Liu, Peng Li, Juanzi Li, et al. 2021. On transferability of prompt tuning for natural language processing. *arXiv preprint arXiv:2111.06719*.

Yang Sun, Nan Yu, and Guohong Fu. 2021. A discourse-aware graph neural network for emotion recognition in multi-party conversation. In *Findings of the Association for Computational Linguistics: EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 16-20 November, 2021*, pages 2949–2958. Association for Computational Linguistics.

Yueqing Sun, Yu Zhang, Le Qi, and Qi Shi. 2022. Tsgp: Two-stage generative prompting for unsupervised commonsense question answering. *arXiv preprint arXiv:2211.13515*.

Mingxing Tan and Quoc V. Le. 2019. Efficientnet: Rethinking model scaling for convolutional neural networks. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 6105–6114. PMLR.

Zhiliang Tian, Yinliang Wang, Yiping Song, Chi Zhang, Dongkyu Lee, Yingxiu Zhao, Dongsheng Li, and Nevin L Zhang. 2022. Empathetic and emotionally positive conversation systems with an emotion-specific query-response memory. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 6364–6376.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *CoRR*, abs/2302.13971.

Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J. Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2019. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 6558–6569. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. 2018. Graph attention networks. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*.

Fanfan Wang, Zixiang Ding, Rui Xia, Zhaoyu Li, and Jianfei Yu. 2021. Multimodal emotion-cause pair extraction in conversations. *CoRR*, abs/2110.08020.

Rui Xia and Zixiang Ding. 2019. Emotion-cause pair extraction: A new task to emotion analysis in texts. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 1003–1012.

Tianbao Xie, Chen Henry Wu, Peng Shi, Ruiqi Zhong, Torsten Scholak, Michihiro Yasunaga, Chien-Sheng Wu, Ming Zhong, Pengcheng Yin, Sida I. Wang, Victor Zhong, Bailin Wang, Chengzu Li, Connor Boyle, Ansong Ni, Ziyu Yao, Dragomir R. Radev, Caiming

Xiong, Lingpeng Kong, Rui Zhang, Noah A. Smith, Luke Zettlemoyer, and Tao Yu. 2022. Unifiedskg: Unifying and multi-tasking structured knowledge grounding with text-to-text language models. *CoRR*, abs/2201.05966.

Xiaocui Yang, Shi Feng, Daling Wang, Sun Qi, Wenfang Wu, Yifei Zhang, Pengfei Hong, and Soujanya Poria. 2023. Few-shot joint multimodal aspect-sentiment analysis based on generative multimodal prompt. *arXiv preprint arXiv:2305.10169*.

Hanlei Zhang, Hua Xu, and Ting-En Lin. 2021a. Deep open intent classification with adaptive decision boundary. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(16):14374–1438.

Hanlei Zhang, Hua Xu, Ting-En Lin, and Rui Lyu. 2021b. Discovering new intents with deep aligned clustering. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(16):14365–14373.

Zhengkun Zhang, Xiaojun Meng, Yasheng Wang, Xin Jiang, Qun Liu, and Zhenglu Yang. 2022. Unims: A unified framework for multimodal summarization with knowledge distillation. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*, pages 11757–11764.

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.

Xiaopeng Zheng, Zhiyue Liu, Zizhen Zhang, Zhaoyang Wang, and Jiahai Wang. 2022. Ueca-prompt: Universal prompt for emotion cause analysis. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 7031–7041.

Lixing Zhu, Gabriele Pergola, Lin Gui, Deyu Zhou, and Yulan He. 2021. Topic-driven and knowledge-aware transformer for dialogue emotion detection. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 1571–1582.

Zhihong Zhu, Xuxin Cheng, Guimin Hu, Yaowei Li, Zhiqi Huang, and Yuexian Zou. 2024. Towards multimodal sarcasm detection via disentangled multi-grained multi-modal distilling. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 16581–16591.

		Cause Recognition			Pair Extraction			
		P	R	F1	P	R	F1	WF1
Task	- w/o MECPE	56.16	54.39	54.64	48.35	58.68	52.41	60.78
	-w/o MCP	56.24	56.28	56.75	46.16	56.57	53.45	61.63
UPL	-w/o A,T	56.25	56.41	56.09	46.09	56.47	53.72	61.41
	-w/o A,V	58.39	58.54	58.51	48.47	58.36	53.09	62.53
	-w/o T,V	56.43	56.77	56.25	46.14	56.54	53.82	61.29
BPL	-w/o A	59.21	59.47	59.61	48.38	59.06	54.64	63.57
	-w/o V	59.46	59.63	59.62	48.54	58.32	54.07	63.75
	-w/o T	56.57	56.68	56.42	46.22	56.10	53.44	61.63
Context	-w/o THC	57.32	56.36	55.19	47.41	57.26	53.43	62.94
	-w/o hierarchy	58.16	58.33	58.37	47.65	57.48	54.52	62.52
	-w/o |w|	56.61	56.63	56.56	46.63	56.41	52.47	62.43

Table 7: Ablation study of UniMEEC on ECF dataset on cause recognition and pair extraction. T, V and A represent textual, visual and acoustic modalities, respectively. UPL and BPL denotes unimodal and bimodal causal prompts, respectively. Hierarchy denotes the hierarchical structure of THC.