

From Extraction to Generation: Multimodal Emotion-Cause Pair Generation in Conversations

Heqing Ma¹, Jianfei Yu¹, Fanfan Wang¹, Hanyu Cao, and Rui Xia¹

Abstract—As an important task in emotion analysis, Multimodal Emotion-Cause Pair Extraction in conversations (MECPE) aims to extract all the emotion-cause utterance pairs from a conversation. However, there are two shortcomings in the MECPE task: 1) it ignores emotion utterances whose causes cannot be located in the conversation but require contextualized inference; 2) it fails to locate the exact causes that occur in vision or audio modalities beyond text. To address these issues, in this paper, we introduce a new task named Multimodal Emotion-Cause Pair Generation in Conversations (MECPG), which aims to identify the emotion utterances with their emotion categories and generate their corresponding causes in a conversation. To tackle the MECPG task, we construct a dataset based on a benchmark corpus for MECPE. We further propose a generative framework named MONICA, which jointly performs emotion recognition and emotion cause generation with a sequence-to-sequence model. Experiments on our annotated dataset show the superiority of MONICA over several competitive systems. Our dataset and source codes will be publicly released.

Index Terms—Multi-task learning, multimodal emotion cause generation, multimodal emotion recognition.

I. INTRODUCTION

AS A key component of human-like artificial intelligence, emotion analysis has remained an active research topic for decades [1], [2], [3]. With the recent proliferation of conversational data on social media platforms such as YouTube, Multimodal Emotion Analysis in Conversations (MEAC) has gained considerable attention [4] due to its wide application in many challenging tasks such as customer support, social interaction, and automatic psychotherapy. Unlike the traditional emotion recognition task that focuses on understanding the human emotional state of a monologue, analyzing emotions for each utterance in a conversation provides a more accurate and nuanced understanding of each speaker's states and intentions. Moreover, multiple modalities are crucial for MEAC because each modality provides unique and complementary information. Textual content conveys semantic meaning, while audio cues such as tone, pitch, and volume offer insights into the speaker's emotional state. Visual signals like facial expressions and body

language further enrich the emotional context. Integrating these modalities allows for a more comprehensive and accurate understanding of emotions, addressing the limitations of relying on a single modality.

Most existing studies on MEAC primarily focus on recognizing the emotion category expressed in each utterance of a conversation [5], [7], [8], [12], [13]. However, emotion recognition can solely help understand the mental state of each speaker but fails to capture the cause of an individual's emotional state. Since recognizing emotions and their causes in conversations is crucial for many downstream applications such as empathic chatbots and mental health care, Wang et al. [14] recently introduced a task named Multimodal Emotion-Cause Pair Extraction in conversations (MECPE), aiming to extract all potential pairs of emotions and their corresponding causes from a conversation using language, audio, and vision modalities. For instance, in Fig. 1(b), Chandler's *anger* emotion in utterance 3 (u_3 for short) is attributed to others opposing his smoking in u_1 , Rachel's subjective criticism in u_2 , and Chandler's perception of unfairness in u_3 . Since MECPE is defined as an extraction task at the utterance level, it is expected to extract the *emotion* utterance u_3 along with its *cause* utterances u_1 , u_2 , and u_3 as a pair, i.e., (u_3 -*anger*- u_1, u_2, u_3).

However, the MECPE task still faces significant limitations in extracting implicit causes.¹ First, for emotion utterances whose causes cannot be pinpointed to specific utterances in the conversation, the ECF dataset for MECPE does not provide cause annotations. These latent causes, although not explicitly expressed in the conversation, can be inferred by understanding the entire conversation. For example, in Fig. 1(a), the MECPE task fails to locate the cause utterance that triggers the *surprise* emotion of Alan in u_1 , and thus ignores the emotion-category-cause triplet of u_1 . Second, in many cases, the emotion causes occur in the vision or audio modalities, but MECPE solely extracts emotion causes at the utterance level, failing to pinpoint the exact cause clues reflected in the video frames. For instance, in Fig. 1(b), the event that triggers the *anger* emotion of all speakers is the *smoking behavior of Chandler*, which is only reflected in the vision modality. In this case, extracting (u_1 -*anger*- u_1) as the emotion-category-cause triplet in the MECPE task is too coarse-grained to identify the real cause of the *anger* emotion.

¹Guo et al. [55] defined that implicit causes are those not directly mentioned in the text but understood through commonsense reasoning over text semantics. We further extend this definition to include causes that can be inferred through commonsense reasoning over textual contexts, as well as audio and visual clues

Manuscript received 1 February 2024; revised 4 August 2024; accepted 5 August 2024. Date of publication 20 August 2024; date of current version 27 May 2025. This work was supported by the Natural Science Foundation of China under Grant 62076133, Grant 62006117, and Grant 62272232. Recommended for acceptance by A. Etemad. (Corresponding authors: Jianfei Yu; Rui Xia.)

The authors are with the School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing, Jiangsu 210094, China (e-mail: hqma@njust.edu.cn; jfyu@njust.edu.cn; ffwang@njust.edu.cn; hycuo@njust.edu.cn; rxia@njust.edu.cn).

Digital Object Identifier 10.1109/TAFFC.2024.3446646

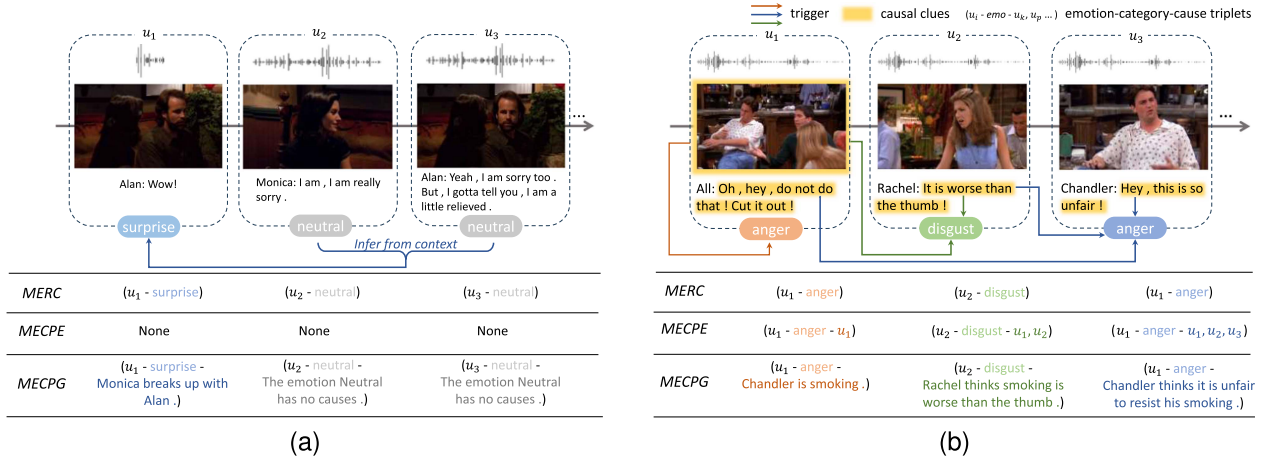


Fig. 1. Comparison among MERC, MECPE, and the proposed MECPG task. MERC identifies and categorizes emotions from utterances without associating them with any causes. MECPE pairs emotion utterances with their corresponding cause utterances but outputs *None* when it cannot pinpoint a specific cause utterance. In contrast, MECPG generates a textual description of causes for each emotion utterance. In (a), the cause of Alan's *surprise* emotion in u_1 is inferred from the whole conversation. In (b), the trigger of friends' *anger* emotion exists in the vision modality, i.e., Chandler's smoking behavior.

To address the aforementioned limitations of the MECPE task, we introduce a new task named Multimodal Emotion-Cause Pair Generation in Conversations (MECPG), which aims to simultaneously recognize emotion utterances with their emotion categories and generate their corresponding causes. For example, in Fig. 1(a), although the MECPE task cannot locate the cause of u_1 , a MECPG system is expected to infer from the awkward atmosphere between Monica and Alan that the event triggering Alan's *surprise* is *Monica breaks up with Alan*. In Fig. 1(b), unlike MECPE that outputs u_1 as the cause of u_1 , our MECPG task aims to explicitly generate the direct cause of u_1 , i.e., *Chandler is smoking*.

To tackle the MECPG task, we construct a dataset named Emotion-Cause-Generation-in-Friends (ECGF) based on a MECPE corpus collected from the sitcom *Friends*, in which we employ annotators to manually annotate the causes for each emotion labeled in the MECPE corpus. We then benchmark the MECPG task with several baseline systems, which combine several representative methods for Multimodal Emotion Recognition and a multimodal generative approach for Emotion Cause Generation in a pipeline manner. Furthermore, we propose a joint framework for Multimodal emOtion recogNI-tion and Cause generAtion (MONICA), which formulates both emotion recognition and cause generation tasks as generation problems and utilizes a pre-trained sequence-to-sequence model BART [15] to simultaneously generate the emotion and its corresponding cause for each utterance in a conversation.

Our main contributions can be summarized as follows:

- We introduce a new task named Multimodal Emotion-Cause Pair Generation in Conversations (MECPG), which aims to jointly identify the emotion utterances with emotion categories and generate their corresponding causes.
- We construct a dataset for the MECPG task based on an existing MECPE corpus, and further propose a joint framework MONICA to generate the emotion category and its corresponding cause for each utterance with a sequence-to-sequence model.

- Experimental results on our annotated dataset show that our proposed framework consistently outperforms many baseline systems in both emotion recognition and cause generation tasks. Further in-depth analysis demonstrates the usefulness of multimodal information in the MECPG task and highlights the advantage of emotion cause generation over emotion cause extraction.

II. DATASET CONSTRUCTION

A. Data Sources

Since there is no available dataset for the MECPG task, we construct our dataset based on the ECF dataset which was recently introduced by [14] for the MECPE task. The ECF dataset consists of video clips selected from the source episodes in popular American sitcom *Friends*, which were originally collected by the MELD dataset [8]. Since the ECF dataset has annotated the emotion category of each utterance and their corresponding utterance-level and span-level causes, in this work, we adopt their emotion annotation and focus on re-annotating the emotion causes by human experts.

B. Annotation Procedure

To ensure consistent and reliable annotations, we recruited two graduate students who are familiar with multimodal emotion analysis to write the emotion cause for utterances with non-neutral emotions, and another three graduate students to evaluate the quality of the annotations.

To assist the two annotators in writing coherent and accurate emotion causes, we provided the following annotation guidelines before annotation:

- Ensure that the annotated sentences are clear and logical, which can accurately reflect the causes of emotions.
- Maintain consistency with the original conversation and restore specific subjects, objects, and referents in final annotations if necessary. Retain the informal language used

TABLE I
COMPARISON BETWEEN OUR ECGF DATASET AND EXISTING EMOTION CAUSE ANALYSIS DATASETS

Dataset	Modality			Conversational	Cause type	#Conversations/Posts	#Emotion utterances	#Avg. cause tokens
	A	V	T					
RECCON-DD [20]	✗	✗	✓	✓	extractive	1106	5861	15.74
RECCON-IE [20]	✗	✗	✓	✓	extractive	16	523	24.06
COVIDET [21]	✗	✗	✓	✗	abstractive	1485	-	26.9
ConvECPE [23]	✗	✗	✓	✓	extractive	151	5725	-
ECF [14]	✓	✓	✓	✓	extractive	1374	7079	12.5
ECGF (Ours)	✓	✓	✓	✓	abstractive	1374	7690	9.9

in the original text or introduce new synonymous phrases to enhance the diversity and readability of the causes.

- Implicit causes are causes that are not explicitly stated or directly mentioned but can be inferred from multimodal context, behavior, or other indirect evidence. These causes often lie beneath the surface and require analysis, reasoning, or background knowledge to uncover.
- Annotators are allowed to revise previous annotations during the annotation process to address any potential issues or discrepancies, ensuring the quality and consistency of the annotations.

We then asked each annotator to independently write the emotion causes of each utterance in 100 conversations after watching the complete video clips. Once their annotations were thoroughly reviewed and passed the quality check, they could proceed to annotate the whole dataset.

Next, we randomly shuffled the 1374 conversations in the ECF dataset and evenly distributed them between the two annotators. Each conversation was annotated by one annotator. First, the annotator needs to watch a complete video of the conversation, which contains textual, acoustic, and visual information, to develop a comprehensive understanding of the background, context, and interaction between speakers. Next, we provided relevant information for each utterance, such as the timestamps, speaker, and emotion category. Based on the video they watched, the annotators were asked to concisely summarize the causes behind specific emotions, consisting of one to three sentences. For utterances with a neutral emotion, annotators were instructed to write “The emotion Neutral has no causes”. In cases where the causes could not be directly inferred from the video, annotators were allowed to make reasonable inferences, as long as they are logical and consistent with the facts presented in the video. During the annotation stage, we regularly checked the work of annotators and provided timely feedback and guidance.

Finally, the cause annotation of each utterance was reviewed by three annotators, who were instructed to score their acceptability. Each annotator was asked to give a ‘pass’ or ‘fail’ assessment for each cause annotation. If at least two evaluators gave a ‘fail’ assessment, the annotation would be discarded. As shown in Table II, the three evaluators reached a consensus for approving around 88.3% of the annotations, indicating satisfactory performance of our annotators in cause annotation. Both the average Cohen’s Kappa score [50] and the Fleiss’ Kappa [51] exceed 0.6, demonstrating a substantial agreement among the three evaluators. For unacceptable cause annotations, each evaluator was asked to provide specific reasons for rejection

TABLE II
THE EVALUATION AGREEMENT FOR ANNOTATED EMOTION CAUSE

Annotator Pair	A&B	A&C	B&C	All
Passing Rate	-	-	-	0.8830
Cohen’s Kappa	0.5907	0.6809	0.5635	0.6117
Fleiss’ Kappa	0.6123			

A, B, and C represent the three evaluators, respectively.

(e.g., irrelevant to the conversation or unreasonable cause). We then assigned those conversations with unacceptable cause annotations to another annotator to revise the causes based on the video clips and the feedback from the evaluators. These revisions are regarded as the final annotations of these utterances.

C. Dataset Statistics and Analysis

Comparison with existing related datasets: In Table I, we compare our ECGF dataset with several existing emotion cause analysis datasets. It can be seen that most previous datasets provided extractive causes, which consist of multiple cause utterances or multiple textual spans in the utterances. For a fair comparison, we concatenate their extracted span-level causes to calculate the number of average cause tokens. Compared to the abstractive causes in our dataset, the emotion causes in extractive datasets tend to be longer and more convoluted. Unlike COVIDET which summarizes the emotion causes for each textual post, our dataset summarizes the causes for each emotion utterance in multimodal conversations. Furthermore, we want to point out that a subset of ECGF was utilized in our recent conference paper [64]. However, there are two major differences between [64] and this work: (1) The two works target different tasks. Wang et al. [64] focused only on the part of the data where non-neutral emotions are annotated, generating causes for given emotions. In contrast, this paper considers the entire conversation, not only recognizing emotions of all utterances but also generating causes, including utterances with neutral emotions. (2) This paper provides a detailed description of the data annotation process and conducts an in-depth statistical analysis of our dataset.

Emotion and Cause Distribution: We selected 1,374 conversations containing 13,619 utterances from the ECF dataset, of which 7,690 utterances are annotated with one of the six emotion categories. Note that there are 932 emotion utterances in the ECF dataset that do not have emotion causes, whereas our annotators have summarized the emotion cause for these utterances. Moreover, Fig. 2 presents the distribution of emotion categories

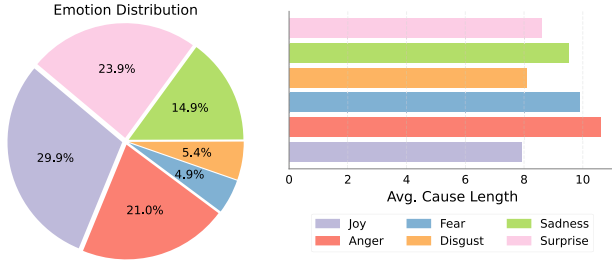


Fig. 2. The distribution of non-neutral emotion utterances and the average length of their corresponding causes.

TABLE III
COMPARISON BETWEEN OUR ANNOTATED CAUSES AND ORIGINAL SPAN-LEVEL CAUSES IN THE ECGF DATASET

u₁: Phoebe (anger): I, umm, shut up!
u₂: Phoebe (surprise): "Good... bye Phoebe and Ursula. I will miss you. P. S. Your Mom lives in Montauk." You just wrote this!
u₃: Ursula (neutral): Well, it is pretty much the gist. Well, except for the poem. You read the poem, right?
u₄: Phoebe (surprise): Noooo!!
u₅: Ursula (neutral): All right, hang on!
Emotion utterance: u ₁
Span-level cause: None
Annotated cause: Ursula didn't tell Phoebe about their birth mother. (Inferred Causes)
Emotion utterance: u ₂
Span-level cause: "You just wrote this"
Annotated cause: Ursula just wrote this suicide note. (Rewritten Causes)
Emotion utterance: u ₄
Span-level cause: "it is pretty much the gist. Well, except for the poem. You read the poem, right?"; "Noooo"
Annotated cause: There was a poem in Phoebe's mother's suicide note. (Conclusive Causes)

and the average length of handwritten causes. This indicates an imbalanced emotion distribution and varying lengths of causes across different emotions. Among all emotions, *Joy* covers the largest percentage and *Fear* covers the smallest. In terms of cause length, *Anger* has the most average cause tokens, while *Joy* has the least. This could be because causes that trigger the *Joy* emotion are often easier to summarize, whereas events that evoke anger may require more specific and detailed descriptions.

Annotation Analysis: To check the quality of our annotations, we further measured the agreement between our annotated causes and the textual span-level causes annotated by the ECF dataset. Note that if an utterance has more than one span-level cause in ECF, we concatenated them as the entire emotion cause. Next, we used the metrics BLEU4 [18], METEOR [19], and ROUGE_L [16] to measure the grammatical similarity and the ConSERT [17] metric to assess the semantic similarity between our annotation and the span-level annotation in ECF. Due to the extracted span-level causes containing many important factors contributing to the emotions, they are semantically similar to our annotated ones, with a ConSERT score of 45.75%. Meanwhile, the manually rewritten causes have removed unnecessary words and replaced unclear expressions with explicit words or phrases outside the conversation, resulting in lower lexical overlap, with BLEU4, METEOR, and ROUGE_L scores of 20.32%, 21.07%, and 33.52%, respectively. Table III presents an example from

our dataset, comparing our annotated causes and the original span-level causes.

III. METHODS

In this section, we detail the proposed joint framework of Multimodal emOtion recogNItion and Cause generAtion (MONICA).

As shown in Fig. 3, MONICA formulates the Multimodal Emotion Recognition in Conversations (MERC) task and the Cause Generation task as generation problems. It employs a shared BART-based encoder-decoder model [24] to jointly predict the emotion and generate its corresponding cause for each utterance.

A. Task Formulation

Given an input conversation, let us use $U = [u_1, \dots, u_n]$ to denote its n utterances. Each utterance is a multimodal input with text, audio, and vision modalities, denoted by $u_i = [u_i^t, u_i^a, u_i^v]$. The goal of our Multimodal Emotion-Cause Pair Generation in Conversations (MECPG) task is to identify the emotion category of each utterance and generate its corresponding emotion cause:

$$\mathcal{P} = \{\dots, (e_i, c_i), \dots\}, \quad (1)$$

where e_i denotes the emotion category of the i -th utterance u_i , which belongs to six basic emotions defined by [11] (i.e., *surprise*, *joy*, *anger*, *fear*, *sadness*, *disgust*) and *neutral*, and c_i is the generated emotion cause of u_i . In Fig. 1, the last row illustrates the outputs of the MECPG task for each utterance.

B. Unimodal Feature Representation

We first obtain the feature representation of three modalities for each utterance u_i as follows.

Text Representation: The textual input of u_i is fed to the word embedding matrix of BART to obtain the text representation $\mathbf{E}_i \in \mathbb{R}^{d \times m}$, where d and m are the hidden size and the length of the textual input.

Audio Representation: We employ HuBERT [56], a self-supervised speech representation learning model, to extract 1024-dimensional audio features from each utterance. To project it into the same dimension of text, we stack a linear layer to obtain $\mathbf{a}_i \in \mathbb{R}^d$.

Visual Representation: Following [14], we apply C3D [32], one kind of 3D-CNN, to extract the 128-dimensional visual features from the video of each utterance. Specifically, we split each video into 16 frames with a resolution of 171×128 , and then feed the 16 frames into the C3D network to obtain a 4096-dimensional visual feature vector. We then add a linear layer to transform it to the dimension of text, denoted by $\mathbf{v}_i \in \mathbb{R}^d$.

C. Multimodal Emotion Recognition in Conversations (MERC)

Since many pre-trained generative models such as BART have been shown to achieve success in both classification and generation tasks [24], we use BART as the backbone of our framework to perform both MERC and Cause generation.

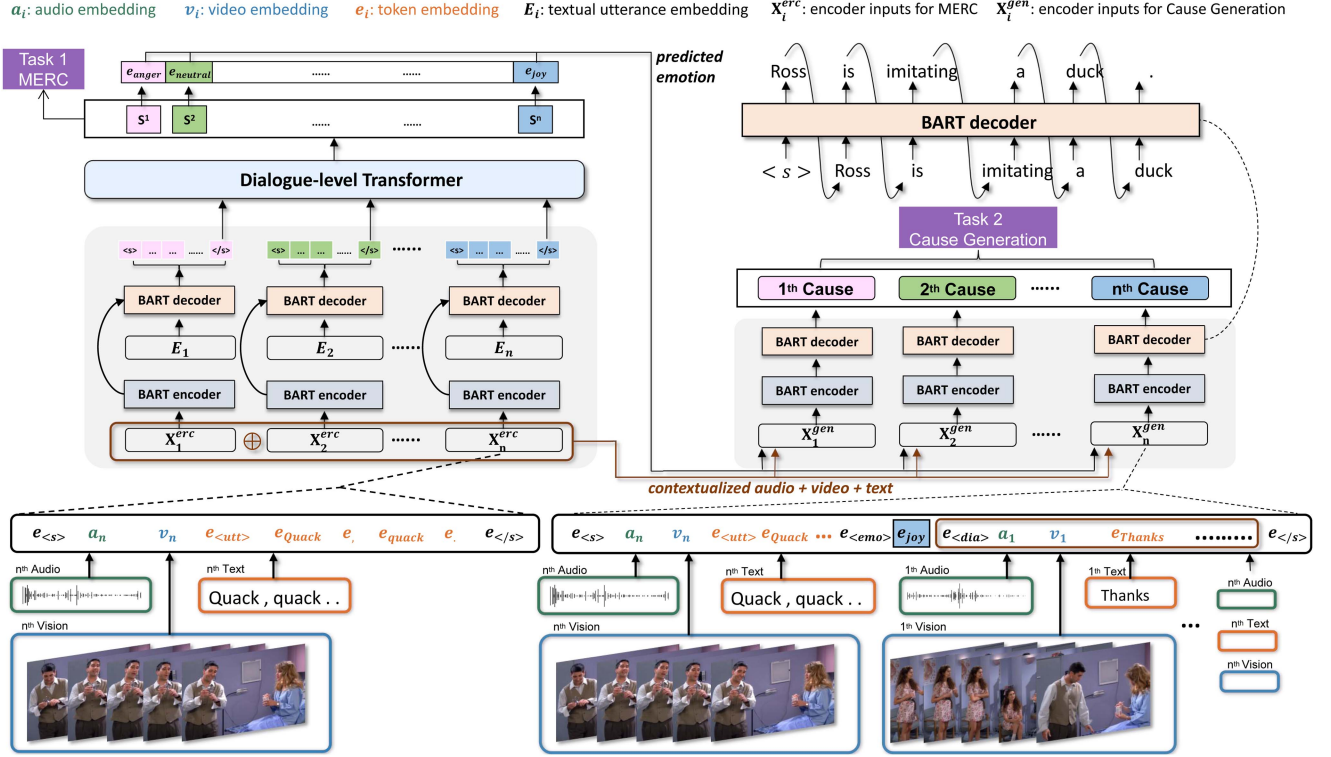


Fig. 3. Overview of our proposed Multimodal emOtion recogNition and Cause generAtion framework (MONICA). Initially, audio feature \mathbf{a}_i and video feature \mathbf{v}_i are concatenated to the textual representation \mathbf{E}_i of each utterance within the conversation to obtain the combined embedding \mathbf{X}_i^{erc} . \mathbf{X}_i^{erc} and \mathbf{E}_i are then fed into the encoder and decoder respectively to get the hidden states of the last token $\langle /s \rangle$. These features are subsequently passed into a dialogue-level transformer to facilitate context interaction for MERC. For Cause Generation, the input \mathbf{X}_i^{gen} is concatenated with the predicted emotion token e_{emo} and the contextualized representation of the window-sized conversation. The decoder then autoregressively generates the cause for each utterance.

First, we design a hierarchical architecture for MERC, which feeds each utterance to BART to obtain the utterance representation, followed by a dialogue-level Transformer to model the inter-utterance interaction for emotion prediction.

Specifically, for each utterance, we concatenate the representations of three modalities as follows:

$$\mathbf{X}_i^{erc} = [\mathbf{e}_{\langle s \rangle}, \mathbf{a}_i, \mathbf{v}_i, \mathbf{e}_{\langle utt \rangle}, \mathbf{E}_i, \mathbf{e}_{\langle /s \rangle}] \quad (2)$$

where $\mathbf{e}_{\langle s \rangle}$ and $\mathbf{e}_{\langle /s \rangle}$ are the embeddings of two special tokens to indicate the beginning and the end of the utterance, and $\mathbf{e}_{\langle utt \rangle}$ serves as a marker to indicate the beginning of textual input.

The multimodal representation \mathbf{X}_i^{erc} is then fed into the BART encoder to obtain its hidden representation. Next, following the practice of applying BART to classification tasks [24], we feed the text representation of the input utterance $\mathbf{E}_i \in \mathbb{R}^{d \times m}$ to the BART decoder and regard the representation of the end token $\langle /s \rangle$ in the last decoder layer as the aggregated representation of the i -th utterance:

$$\begin{aligned} \mathbf{H}_i^{erc} &= \text{BART-Encoder}(\mathbf{X}_i^{erc}), \\ \mathbf{h}_i^{erc} &= \text{BART-Decoder}(\mathbf{H}_i^{erc}; \mathbf{e}_{i, < t \rangle}), \\ \hat{\mathbf{h}}_i &= \mathbf{h}_i^{erc}, \end{aligned} \quad (3)$$

where $\mathbf{e}_{i, < t \rangle}$ denotes the embeddings of the first $t - 1$ tokens in the text input of the i -th utterance.

We further employ a dialogue-level Transformer [25] as the context encoder in (4). The multi-head attention mechanism, by considering different perspectives simultaneously during attention computation, can capture diverse and complex patterns within the conversation. This enables a more comprehensive understanding of the interactions between utterances and facilitates a richer representation of the conversation.

$$\text{Att}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{Softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}}\right)\mathbf{V},$$

$$\text{head}_i = \text{Att}\left(\hat{\mathbf{h}}_j \mathbf{W}_i^Q, \hat{\mathbf{h}}_k \mathbf{W}_i^K, \hat{\mathbf{h}}_l \mathbf{W}_i^V\right),$$

$$\text{MHead}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{Concat}(\text{head}_1, \text{head}_2, \dots, \text{head}_l) \mathbf{W}^O,$$

where $\hat{\mathbf{h}}_j$ and $\hat{\mathbf{h}}_k$ represent any two utterances, $\mathbf{W}_i^Q \in \mathbb{R}^{d \times d_q}$, $\mathbf{W}_i^K \in \mathbb{R}^{d \times d_k}$, $\mathbf{W}_i^V \in \mathbb{R}^{d \times d_v}$, and $\mathbf{W}^O \in \mathbb{R}^{d \times d}$ are learnable parameters, d_q , d_k , and d_v are dimensions of query, key, and value matrices (i.e., \mathbf{Q} , \mathbf{K} , and \mathbf{V}), and l indicates the number of attention heads. With the dialogue-level Transformer, we obtain the context-aware representation of each utterance \mathbf{h}_i as follows:

$$[\mathbf{h}_1, \dots, \mathbf{h}_n] = \text{Dialogue-Transformer}\left([\hat{\mathbf{h}}_1, \dots, \hat{\mathbf{h}}_n]\right), \quad (4)$$

Subsequently, \mathbf{h}_i is fed into a classifier to predict the corresponding emotion label \hat{e}_i .

$$\mathbf{P}_i = \text{Softmax}(\mathbf{W}^T \mathbf{h}_i + \mathbf{b}),$$

$$\hat{e}_i = \operatorname{argmax}(\mathbf{P}_i), \quad (5)$$

where $\mathbf{W} \in \mathbb{R}^{d \times K}$, $\mathbf{b} \in \mathbb{R}^K$ are weight parameters to learn, and $K=7$ denotes the number of emotion classes.

Finally, we use the cross-entropy loss to optimize the parameters for the MERC task:

$$\mathcal{L}_{erc} = -\frac{1}{n} \sum_{i=1}^n \sum_{j=1}^K e_{ij} \cdot \log \hat{e}_{ij}, \quad (6)$$

D. Cause Generation

For cause generation, if the emotion category of an input utterance is one of the six emotion categories, we expect our model to generate the annotated cause; otherwise, it is expected to generate “*The Emotion Neutral has no causes*”. Since it is important to consider both the input utterance and its conversational context for emotion cause generation, we concatenate the representation of the current utterance \mathbf{E}_i and its context as the input below:

$$\begin{aligned} \mathbf{D}_i^{win} &= [\mathbf{e}_{(dia)}, \mathbf{a}_{i-win}, \mathbf{v}_{i-win}, \mathbf{E}_{i-win}, \dots, \mathbf{a}_{i+win}, \\ &\quad \mathbf{v}_{i+win}, \mathbf{E}_{i+win}], \\ \mathbf{X}_i^{gen} &= [\mathbf{e}_{(s)}, \mathbf{a}_i, \mathbf{v}_i, \mathbf{e}_{(utt)}, \mathbf{E}_i, \mathbf{e}_{(emo)}, \mathbf{e}_{emo}, \mathbf{D}_i^{win}, \mathbf{e}_{(s)}], \end{aligned}$$

where $\mathbf{e}_{(emo)}$ and $\mathbf{e}_{(dia)}$ are special symbols indicating the emotion category and the beginning of the context, and \mathbf{D}_i^{win} denotes the context within the window, i.e., the audio, visual, and textual features of utterances within the range from $i - win$ to $i + win$. Here, win represents the size of the contextual window, determining the number of preceding and succeeding utterances to consider for context. For example, in Fig. 3, the input \mathbf{X}_n^{gen} concatenates the predicted emotion e_{joy} and the multimodal context (highlighted in the brown box).

Next, each concatenated utterance is fed into the BART-Encoder to get their hidden states, and then the BART-Decoder generates the output sequence. Specifically, at each time step t , the BART-Decoder takes the previous predicted tokens $w_{<t}$ and the hidden states of the BART-Encoder \mathbf{H}_i^{gen} as inputs, and then predicts the probability of the next token with a linear layer followed by a Softmax layer below:

$$\begin{aligned} \mathbf{H}_i^{gen} &= \text{BART-Encoder}(\mathbf{X}_i^{gen}), \\ \mathbf{h}_t^{gen} &= \text{BART-Decoder}(\mathbf{H}_i^{gen}; w_{<t}), \\ \mathbf{p}(w_t) &= \text{Softmax}(\mathbf{W}_{gen}^\top \mathbf{h}_t^{gen} + \mathbf{b}_{gen}), \end{aligned} \quad (7)$$

where w_t is the token generated at time step t , and $w_{<t}$ represents the sequence of previously generated tokens. Here, $\mathbf{W}_{gen} \in \mathbb{R}^{d \times |V|}$ and $\mathbf{b}_{gen} \in \mathbb{R}^{|V|}$ are learnable parameters, where $|V|$ is the size of the vocabulary.

During training, we adopt the teacher-forcing strategy to train our model and use the negative log-likelihood (i.e., cross-entropy) loss to optimize parameters for the cause generation task:

$$\mathcal{L}_{gen} = -\sum_{t=1}^m \sum_{j=1}^{|V|} e_{tj} \log \hat{e}_{tj}, \quad (8)$$

where m is the length of the cause and θ is the parameters of BART need to be optimized. During inference, we replace the true emotion label with the predicted one for cause generation. Moreover, we use beam search to obtain more reasonable causes.

E. Training Mechanism

Since our framework consists of two components that solve two related tasks, we design the following two training approaches:

Pipeline Training: We first train on the MERC task and then perform cause generation according to the predicted emotions from the first step. The training objectives are given in (6) and (8), respectively.

Joint Training: Another strategy is to merge the two components by simultaneously optimizing the parameters for MERC and cause generation in a multi-task learning manner. We use the weighted sum of \mathcal{L}_{erc} and \mathcal{L}_{gen} as the training objective:

$$\mathcal{L} = (1 - \beta)\mathcal{L}_{erc} + \beta\mathcal{L}_{gen}. \quad (9)$$

IV. EXPERIMENTS

A. Evaluation Metrics

Because the MECPG task involves a classification subtask (MERC) and a generation subtask (Cause generation), we consider using separate evaluation metrics for the two subtasks.

Multimodal Emotion Recognition in Conversations (MERC): Considering the imbalanced distribution of emotion utterances in the ECF dataset, we follow [8] to use Weighted_F1 as the metric, which is the weighted average of F1 scores considering the number of instances for each emotion category. In addition, we also consider Binary_F1 by treating MERC as a binary classification task, in which non-neutral emotions are regarded as positive while neutral ones as negative.

Cause Generation: For cause generation, we employ BLEU4 [18], METEOR [19], ROUGE-L [16] and CIDEr [28] to evaluate the syntactic similarity between the generated causes and the references. We also use Sem-Sim [29], BLEURT [30] and BERTScore [27] to measure the semantic similarity between the generated candidate and the ground truth. To compute BERTScore, the tokenized reference sentence $x = \langle x_1, \dots, x_k \rangle$ and candidate $\hat{x} = \langle \hat{x}_1, \dots, \hat{x}_l \rangle$ are fed into a pretrained RoBERTa-Large model to compute contextual embeddings, i.e., $\langle \mathbf{x}_1, \dots, \mathbf{x}_k \rangle$ and $\langle \hat{\mathbf{x}}_1, \dots, \hat{\mathbf{x}}_l \rangle$. The cosine similarity of a reference token x_i and a candidate token \hat{x}_j is reduced to the inner product $\mathbf{x}_i^\top \hat{\mathbf{x}}_j$. Then the F1 measure F_{BERT} can be calculated as follows:

$$\begin{aligned} R_{\text{BERT}} &= \frac{1}{|x|} \sum_{x_i \in x} \max_{\hat{x}_j \in \hat{x}} \mathbf{x}_i^\top \hat{\mathbf{x}}_j, \\ P_{\text{BERT}} &= \frac{1}{|\hat{x}|} \sum_{\hat{x}_j \in \hat{x}} \max_{x_i \in x} \mathbf{x}_i^\top \hat{\mathbf{x}}_j, \\ F_{\text{BERT}} &= 2 \frac{P_{\text{BERT}} \cdot R_{\text{BERT}}}{P_{\text{BERT}} + R_{\text{BERT}}}. \end{aligned} \quad (10)$$

TABLE IV
PERFORMANCE COMPARISON ON THE MULTIMODAL EMOTION RECOGNITION IN CONVERSATIONS (MERC) SUBTASK

Methods		Binary_F1	Weighted_F1							
			All	Neutral	Joy	Sadness	Anger	Fear	Surprise	Disgust
Pipeline	MMGCN	0.7542	0.5843	0.7378	0.5520	0.3676	0.4785	0.0317	0.5984	0.0247
	MMDFN	0.7662	0.5819	0.7309	0.5477	0.4161	0.4888	0.0000	0.5727	0.0000
	UniMSE	0.7650	0.5845	0.7505	0.5570	0.3423	0.4242	0.0952	0.6087	0.0460
	MECPE_2Steps	0.7899	0.5947	0.7520	0.5738	0.3511	0.4678	0.0000	0.6399	0.0000
	MONICA_T5 (Pipeline)	0.7610	0.5851	0.7315	0.5514	0.3820	0.4430	0.1455	0.5813	0.2353
	MONICA_BART (Pipeline)	0.7926	0.5918	0.7391	0.6027	0.3816	0.4479	0.0625	0.6027	0.0222
Joint	MMT5 (Joint)	0.7520	0.5660	0.7137	0.5413	0.3090	0.4869	0.0000	0.5980	0.0000
	MMBART (Joint)	0.7505	0.5807	0.7219	0.5415	0.3494	0.4786	0.1408	0.6009	0.1579
	MONICA_T5 (Joint)	0.7876	0.5922	0.7494	0.5721	0.3834	0.4486	0.0678	0.5852	0.1111
	MONICA_BART (Joint)	0.7997	0.6107	0.7640	0.6120	0.4047	0.4727	0.0635	0.6178	0.0000

In line with the process of the MELD dataset in [34], uniMSE only used audio features.

B. Experimental Settings

We divide our ECGF dataset into training, validation, and test sets at the conversation level based on a 7:1:2 ratio. During the training stage, we tune the hyper-parameters on the validation set. After tuning, we set the learning rate to $1e-5$ and the batch size to 2. We utilize Adam as the optimizer and set weight decay to 0.01. The trade-off parameter β for the multi-task learning loss is set to 0.9. The maximum length for an individual utterance is 64 and the maximum context length is 400. The generated causes are limited to no more than 60 tokens.

C. Baseline Systems

Since MECPG is a new task, we first consider comparing our method with several pipeline baselines, which first employ a representative method for MERC to predict the emotion category of each utterance, and then use a multimodal generative approach to generate their corresponding causes as follows: (1) **MMGCN+MMBART** first adopts a multimodal fusion graph convolutional network [33] to capture the multimodal and long-range contextual information for emotion prediction. Next, it utilizes BART to generate causes based on multimodal inputs as described in Section III-D, and we name this step as MMBART. (2) **UniMSE+MMBART** replaces the MERC model in MMGCN+MMBART with a multimodal sentiment knowledge sharing framework named UniMSE [34], which performs multimodal fusion at both syntactic and semantic levels and incorporates contrastive learning between modalities and samples to capture the differences and consistencies within various emotion categories. (3) **MMDFN+MMBART** replaces the MERC model with a multimodal dynamic fusion network named MM-DFN [35], which aims to identify the emotion of each utterance by utilizing a graph-based dynamic fusion module to integrate multimodal contextual features in the conversation. (4) To compare the performance with pre-trained models of similar size, we also replace BART with T5 in those pipelines, i.e., **MMGCN+MMT5**, **UniMSE+MMT5** and **MMDFN+MMT5**.

Considering the recent trend of applying Large Language Models (LLMs) to various NLP tasks, we apply GPT3.5 [36] to cause generation in a 5-shot learning setting and yield four pipelines: **MMGCN+GPT3.5**, **UniMSE+GPT3.5**, **MMDFN+GPT3.5** and **MECPE_2Steps+GPT3.5** which employs the method introduced by Wang et al. [14] for emotion

TABLE V
THE PROMPT TEMPLATE AND FEW-SHOT EXAMPLES FOR GPT3.5 TO GENERATE EMOTION CAUSES

Instruction

Write a sentence to conclude the cause of the utterance's emotion.
I will give you the whole conversational context to help write the cause.
The cause should be clear and short and you can copy the original expressions in the conversational context.
Maybe some causes need your imagination.
The cause must strictly follow the fixed prefix: 'The cause is'
The emotion's categories are {neutral, surprise, anger, joy, disgust, fear, sadness}.
Especially, when the emotion is neutral, write 'The emotion Neutral has no causes' as cause.
The target utterance is also from the context.
The context is composed of multi-turn utterances, each utterance is composed of the speaker and his/her words.
The cause's length should not exceed 60 tokens.
I will give a few examples for you.
Remember, don't ask me any questions or seek any other information.

Few-shot Examples

Here are five examples:
utterance: Well , what ? What ? What is it ? That she left you ? That she likes women ? That she left you for another woman that likes women ? // emotion: anger // context: Chandler : Well , what ? What ? What is it ? That she left you ? That she likes women ? That she left you for another woman that likes women ? Ross : Little louder , okay , I think there is a man on the twelfth floor in a coma that did not quite hear you ... Chandler : Then what ? Ross : My first time with Carol was ... Joey : What ? Ross : It was my first time . Joey : With Carol ? Oh . Chandler : So in your whole life , you have only been with one ... oh . Joey : Whoah , boy , hockey was a big mistake ! There was a whole bunch of stuff we could have done tonight !
The cause is "Ross's wife left him because she likes women ."
(Other examples)

utterance extraction and GPT-3.5 for cause generation. The prompt template for GPT-3.5 is shown in Table V.

We further propose several joint methods for fair comparison: **MMBART (Joint)** is a joint baseline that employs the MMBART architecture to simultaneously generate the emotion predictions and their corresponding causes in the decoder. **MMT5 (Joint)** and **MONICA_T5 (Joint)** replace the backbones of MMBART (Joint) and our **MONICA_BART (Joint)** with T5, respectively.

D. Main Results

a) *Automatic Evaluation:* In Table IV, we report the results of our MONICA framework and all the baseline systems on the MERC subtask. Due to the imbalanced emotion distribution in ECGF, all the models exhibit significant variations in predicting different emotions. In the test set, the number of ground-truth

TABLE VI
PERFORMANCE COMPARISON ON THE CAUSE GENERATION SUBTASK WITH RESPECT TO DIFFERENT NUMBER OF EMOTION CATEGORIES

Methods		Seven Emotion Categories						Six Emotion Categories (Excluding Neutral)							
		BLEU4	METEOR	ROUGE	CIDEr	Sem-Sim	BLEURT	F _{bert}	BLEU4	METEOR	ROUGE	CIDEr	Sem-Sim	BLEURT	F _{bert}
LLM	MMGCN+GPT3.5	0.1980	0.2402	0.3517	0.7654	0.6707	-0.1493	0.9068	0.0448	0.1359	0.1519	0.4000	0.5367	-0.6921	0.8707
	MMDFN+GPT3.5	0.1899	0.2370	0.3411	0.7427	0.6682	-0.1573	0.8938	0.0466	0.1431	0.1580	0.4204	0.5575	-0.6507	0.8601
	UniMSE+GPT3.5	0.1987	0.2446	0.3570	0.7510	0.6818	-0.1230	0.9083	0.0380	0.1364	0.1502	0.3642	0.5440	-0.6809	0.8710
	MECPE_2Steps+GPT3.5	0.1971	0.2442	0.3531	0.7592	0.6818	-0.1299	0.9077	0.0463	0.1444	0.1588	0.4121	0.5606	-0.6503	0.8732
Pipeline	MMGCN+MMT5	0.3829	0.2769	0.5014	4.3573	0.6570	-0.1879	0.9287	0.1889	0.1753	0.2694	1.5520	0.5136	-0.7320	0.8961
	MMDFN+MMT5	0.3848	0.2773	0.4971	4.2998	0.6559	-0.1944	0.9280	0.2051	0.1832	0.2805	1.6362	0.5261	-0.7064	0.8975
	UniMSE+MMT5	0.3857	0.2787	0.5045	4.3790	0.6596	-0.1863	0.9291	0.1937	0.1774	0.2742	1.5799	0.5169	-0.7355	0.8966
	MMGCN+MMBART	0.3985	0.2876	0.5201	4.5105	0.6733	-0.1597	0.9309	0.2012	0.1843	0.2866	1.6666	0.5311	-0.7129	0.8978
	MMDFN+MMBART	0.3940	0.2862	0.5125	4.4026	0.6695	-0.1769	0.9298	0.2153	0.1944	0.3030	1.7793	0.5476	-0.6852	0.9004
	UniMSE+MMBART	0.4007	0.2891	0.5262	4.5790	0.6791	-0.1476	0.9320	0.1962	0.1816	0.2859	1.6645	0.5311	-0.7190	0.8979
	MONICA_T5 (Pipeline)	0.3827	0.2773	0.5010	4.3411	0.6570	-0.1886	0.9285	0.1932	0.1786	0.2767	1.5992	0.5191	-0.7198	0.8969
	MONICA_BART (Pipeline)	0.3967	0.2881	0.5090	4.3372	0.6713	-0.1815	0.9297	0.2356	0.2060	0.3203	1.8971	0.5677	-0.6478	0.9037
Joint	MMT5 (Joint)	0.3886	0.2830	0.5125	4.2865	0.6716	-0.1460	0.9299	0.1998	0.1906	0.3187	1.7216	0.5591	-0.6023	0.9022
	MMBART (Joint)	0.4003	0.2885	0.5720	4.4218	0.6696	-0.1723	0.9294	0.2244	0.1963	0.3842	1.7855	0.5437	-0.6785	0.8993
	MONICA_T5 (Joint)	0.3858	0.2776	0.4893	4.2979	0.6390	-0.2106	0.9252	0.2283	0.1887	0.2749	1.7334	0.5051	-0.7172	0.8949
	MONICA_BART (Joint)	0.4049	0.2939	0.5271	4.5325	0.6844	-0.1478	0.9320	0.2205	0.1981	0.3099	1.8163	0.5585	-0.6714	0.9015

emotion labels for *joy*, *sadness*, *anger*, *fear*, *surprise*, and *disgust* is 429, 241, 333, 56, 307, and 79, respectively. It is clear to see that the BART-based generative model *MONICA (Pipeline)* performs better than all the other pipeline systems in Binary_F1 and Weighted_F1, which indicates the effectiveness of our generative framework on the MERC task. Moreover, our multi-task learning framework *MONICA (Joint)* further boosts the performance of *MONICA (Pipeline)* with an improvement of 0.71% and 1.89% in Weighted_F1 and Binary_F1, respectively. We conjecture the reason is that the two subtasks are mutually indicative, and the multi-task learning framework enforces the interactions between the two subtasks, consequently improving the performance of the MERC subtask. Compared to Joint approaches, it can be seen that the Pipeline methods are not necessarily worse. Both MMT5 and MMBART show a performance drop compared to the other Pipeline methods, indicating that directly decoding emotions and causes simultaneously from the output sequence is not conducive to the model learning correct emotions.

In addition, we report the performance of each method on the cause generation task in Table VI. For a fair comparison, we divide the test set based on the ground-truth emotion labels of each utterance and provide results under seven emotions and six emotions (excluding neutral). It can be seen that *MONICA (Joint)* performs best considering seven emotions, but *MONICA (Pipeline)* achieves the best average performance across the six emotion categories. We believe that cause generation is much more complex than the emotion recognition task, and striking a balance in the loss function for both tasks is challenging. This leads to the fact that although the joint framework brings a noticeable improvement on the MERC subtask, it may have an adverse impact on cause generation for different emotions. During our experiments, we find that compared to BART, T5 is more sensitive to the weight of the loss in multi-task learning, making it more difficult to balance the classification and generation tasks. Therefore, we chose BART as the backbone of our framework.

b) *Human Evaluation*: In addition to automatic evaluation, we also perform human evaluation on 1000 randomly selected test samples. We employ two human evaluators to rate generated

TABLE VII
HUMAN EVALUATION ON THE GENERATED CAUSE SUMMARY

Methods	Contextuality	Fluency	Relevance
GPT3.5	3.12	4.57	2.66
MONICA (Text)	2.95	4.50	2.42
MONICA (Pipeline)	3.26	4.47	2.60
MONICA (Joint)	3.24	4.38	2.81

causes on a scale of 1-5 in terms of three aspects: *Contextuality*, *Fluency*, and *Relevance*. *Contextuality* measures whether the generated sentence aligns with the context and is coherent with the preceding conversation. *Fluency* assesses the smoothness and naturalness of the sentence in terms of grammar and language usage. *Relevance* is used to judge whether the generated sentence includes the exact cause for the expressed emotion. The average of two evaluators' scores are presented in Table VII. Note that *MONICA (Text)* is the variant of *MONICA (Joint)* that solely utilizes the text modality. Based on the results, we observe that *MONICA_BART (Joint)* obtains a relatively high score in relevance and *MONICA_BART (Pipeline)* achieves best in contextuality, indicating that the causes generated by our framework are generally coherent and relevant to the conversation. Due to the hallucination phenomenon in LLMs, although the causes generated by GPT-3.5 are sufficiently fluent, they introduce some fake information, resulting in reduced contextual consistency.

E. Ablation Study

To investigate the impact of different modalities of our *MONICA (Joint)* approach, we conduct an ablation study. As shown in Table VIII, we explore the fusion of four different audio and visual features, i.e., openSMILE [31], HuBERT, 3DCNN, and ResNet [57]. Clearly, the integration of HuBERT audio features brings the most significant improvement to the model, increasing the Binary_f1 score by 1.02% on the MERC task. It also achieves optimal performance in every metric on the Cause Generation task, demonstrating the good performance of audio features based on the pre-trained language model.

TABLE VIII
ABLATION STUDY OF DIFFERENT MODALITY FEATURES

Methods	Emotion Recognition				Cause Generation				
	Weighted_F1	Binary_F1	BLEU4	METEOR	ROUGE	CIDEr	Sem-Sim	BLEURT	F _{bert}
MONICA_BART (Text)	0.6132	0.7964	0.4017	0.2907	0.5134	4.3847	0.6740	-0.1774	0.9299
+A _{openSMILE}	0.6113	0.7917	0.4001	0.2886	0.5150	4.4229	0.6736	-0.1807	0.9300
+A _{HuBERT}	0.6119	0.8066	0.4063	0.2936	0.5188	4.4686	0.6792	-0.1644	0.9313
+V _{3DCNN}	0.6080	0.8038	0.4036	0.2923	0.5148	4.4104	0.6782	-0.1732	0.9307
+V _{ResNet}	0.6044	0.7941	0.3994	0.2890	0.5099	4.3774	0.6742	-0.1849	0.9296

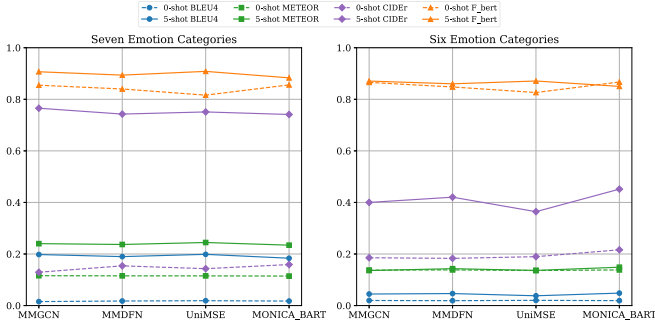


Fig. 4. Performance comparison of GPT-3.5 for cause generation under the 0-shot and 5-shot setting.

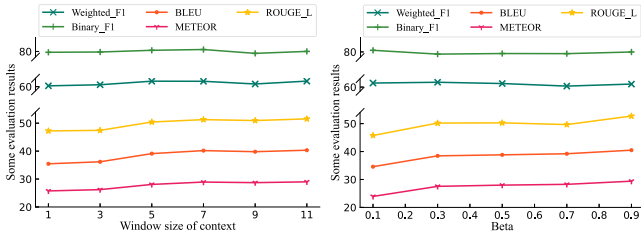


Fig. 5. Impact of the window size and hyper-parameter β .

To explore the impact of prompt samples on the inference ability of LLMs, we present the experimental results of four pipelines with GPT-3.5 under 0-shot and 5-shot settings across four major metrics in Fig. 4. Despite using only five random samples from the validation set, the 5-shot setting results in considerable improvements on both seven and six emotion categories, which reveals that in-context learning can effectively enhance the reasoning capabilities of LLMs.

In addition, we also explore the impact of the trade-off parameter β in multi-task learning and the window size of the context. As shown in Fig. 5, the window size has a much greater impact on cause generation than on emotion recognition, since emotions tend to be expressed in the current utterance. As β increases, the performance of cause generation improves consistently, but MERC exhibits some fluctuations.

F. Case Study

We further conduct a case study to explain the advantages of our MECPG task and our MONICA framework. As shown in Fig. 6, this is a conversation fragment from the ECGF test set. The emotions of u_2 , u_4 , and u_8 are *surprise*, and their

corresponding annotated causes are shown in the last row. For u_2 , the MECPE task annotates u_1 as the cause, which is similar to the expression “Rachel is pregnant” generated by *MONICA (Joint & Pipeline)*. In this case, they are both consistent with the reference. In contrast, GPT3.5 gives a relatively longer cause that contains an unnecessary explanation. For u_8 , the MECPE task annotates both u_7 and u_8 as cause utterances. *MONICA (Pipeline)* captures crucial information but makes incorrect inferences. Both *MONICA (Joint)* and GPT3.5 provide causes that are reasonable but deviate from the ground truth. For the more challenging case like u_4 , which requires the model to exhibit reasoning abilities, our MONICA framework and GPT3.5 generate wrong causes.

V. RELATED WORK

Multimodal Emotion Recognition in Conversations (MERC): ERC is a popular research topic in the field of emotion analysis [52], [53], [54], [58], [59] and MERC extends it to multimodal scenarios. MERC aims to identify the emotions associated with each utterance from a pre-defined set, given the conversation (including text, audio, and video) and information about each speaker [37]. Most existing studies on the MERC task have adopted various neural models to model inter-speaker interactions and capture context information in conversations, including GRU-based methods [38], [60], graph-based methods [6], [6], [33], [35], [40], transformer-based methods [39], [47], data augmentation-based methods [48] and pre-trained language model-based methods [7], [34]. In this work, we follow the last line of work and focus on proposing a BART-based generative model for MERC.

Emotion Cause Analysis: Emotion Cause Analysis (ECA) is an emerging task that aims to analyze underlying emotions and their corresponding causes. It consists of two representative subtasks: Emotion Cause Extraction (ECE) and Emotion Cause Pair Extraction (ECPE). The ECE task was initially introduced by [41] to extract text fragments that indicate the causes of given emotions. By analyzing the corpus proposed by [41], Chen et al. [43] pointed out that clause may be a more suitable unit for cause annotation, they re-defined the ECE task at the clause level and constructed a Chinese emotion cause dataset. Poria et al. [20] extended the task of emotion cause extraction to conversations by introducing a new task that extracts the cause utterances or cause spans corresponding to a given emotion utterance in a dialogue. The RECCON dataset introduced in their work has recently received increasing attention [44], [61], [62].

As ECE requires pre-annotated emotions and ignores the inter-connections between emotion extraction and cause











Conversation Fragment					
u_1		u_2		u_3	
	Monica: She is not pregnant. It is Rachel. Rachel the one who pregnant.		Joey: Oh my God.		Monica: Phoebe I think he would notice if you did not have a baby in nine months!
	u_4		u_5		u_6
					
	Phoebe: It is Joey!		Joey: Now I can not believe it! What? Rachel pregnant? Who the father?		Phoebe: We do not know.
u_7	u_8	u_9	u_{10}		
					
	Joey: Ohh ... I wonder if it is that dude.		Monica: There is a dude?		Joey: Yeah.
					Phoebe: Who? Who is it?
Emotion-Category-Cause Triplets					
MECPE Annotation	$(u_2, \text{surprise}, u_1), (u_4, \text{surprise}, u_3), (u_6, \text{surprise}, u_7), (u_8, \text{surprise}, u_9)$				
MECPG Annotation	$(u_2, \text{surprise}, \text{"Rachel is pregnant."}), (u_4, \text{surprise}, \text{"It is Joey who wants to marry Phoebe."}), (u_6, \text{surprise}, \text{"There is a dude who spent the night with Rachel."})$				
GPT3.5	$(u_2, \text{surprise}, \text{"Joey is surprised to find out that Rachel is pregnant and wonders who the father is."}), (u_4, \text{surprise}, \text{"Phoebe realizes that Joey is the father of Rachel's baby."}), (u_6, \text{surprise}, \text{"Joey mentions that there is a guy."})$				
MONICA (Pipeline)	$(u_2, \text{surprise}, \text{"Rachel is pregnant."}), (u_4, \text{surprise}, \text{"It is Joey's wife who is pregnant."}), (u_6, \text{surprise}, \text{"There is a dude in the room."})$				
MONICA (Joint)	$(u_2, \text{surprise}, \text{"Rachel is pregnant."}), (u_4, \text{surprise}, \text{"It is Joey's wife who is pregnant."}), (u_6, \text{surprise}, \text{"Monica wonders who the father is."})$				

Fig. 6. Comparison between the annotated output of MECPE and MECPG and the generated output of MONICA and GPT3.5.

extraction, Xia and Ding [42] proposed the ECPE task to jointly extract potential emotions and their causes from the text. With the increasing attention on ECA in conversations [9], [10], [20], [44] in recent years, Li et al. [45] introduced the ECPEC task to leverage the emotions of all speakers and understand the mutual influences among utterances, extending the ECPE task from news text to conversational text. Considering that multimodality is especially important for discovering both emotions and their causes in conversations, Wang et al. [14], [63] proposed the MECPE task to explore this task in a multimodal setting, where emotions and causes are annotated based not only on the text, but also the corresponding audio and visual modalities. However, these works mainly focus on emotion cause extraction at the utterance level, whereas we focus on emotion cause generation for each utterance.

Cause Generation: Recent advancements in pre-trained models have shown substantial progress in various inference tasks such as affect-driven dialog generation [65], [66] and causal reasoning in dialogues [22]. Since emotion's triggers always scatter across multiple sentences, some works have recently leveraged text generation methods to generate causes of an emotion expressed in a given text. For example, Riyadh and Shafiq [46] focused on generating meaningful causes for pre-labeled emotions in given sentences, while Zhan et al. [21] attempted to identify emotions and summarize triggering causes in social media.

Nevertheless, these aforementioned works primarily focus on text-only scenarios. Given that conversation in its natural form is multimodal and many emotion causes are triggered by human behavior (i.e., visual modality) or human speech (i.e., audio modality), we aim to explore the multimodal emotion cause generation task in this work.

VI. CONCLUSION

In this paper, we present a new task named Multimodal Emotion-Cause Pair Generation in Conversations (MECPG). It poses a significant challenge as it requires models to possess reasoning abilities while incorporating multimodal information. Moreover, we construct a new dataset named ECGF and develop a joint Multimodal emOtion recognItion and Cause generation (MONICA) framework for our task. Experimental results demonstrate the effectiveness of our MONICA framework on the MECPG task.

Although this work proposes a joint framework to address the MECPG task, it still suffers from some limitations. We observed that for the emotions correctly predicted by MONICA, approximately 18.3% of the generated causes were irrelevant. Further research is needed to improve the performance of joint models and to design fairer evaluation metrics to comprehensively assess both emotions and causes. Additionally, how to better leverage the perceptual capabilities of LLMs to generate more reasonable and diverse causes remains a significant challenge.

ACKNOWLEDGMENTS

The authors would like to thank all anonymous reviewers for their insightful comments.

REFERENCES

- [1] R. W. Picard, *Affective Computing*, MIT press, 2000.
- [2] R. A. Calvo and S. D'Mello, "Affect detection: An interdisciplinary review of models, methods, and their applications," *IEEE Trans. Affect. Comput.*, vol. 1, no. 1, pp. 18–37, Jan. 2010.
- [3] K. Ahmad, *Affective Computing and Sentiment Analysis: Emotion, Metaphor and Terminology*. Berlin, Germany: Springer, 2011.

- [4] S. Poria, D. Hazarika, N. Majumder, and R. Mihalcea, "Beneath the tip of the iceberg: Current challenges and new directions in sentiment analysis research," *IEEE Trans. Affect. Comput.*, vol. 14, no. 1, pp. 108–132, First Quarter 2023.
- [5] F. Chen, Z. Sun, D. Ouyang, X. Liu, and J. Shao, "Learning what and when to drop: Adaptive multimodal and contextual dynamics for emotion recognition in conversation," in *Proc. ACM Int. Conf. Multimed.*, 2021, pp. 1064–1073.
- [6] A. Joshi, A. Bhat, A. Jain, A. Singh, and A. Modi, "COGMEN: Contextualized GNN based multimodal emotion recognition," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics - Hum. Lang. Technol.*, 2022, pp. 4148–4164.
- [7] Y. Mao, G. Liu, X. Wang, W. Gao, and X. Li, "DialogueTRM: Exploring multi-modal emotional dynamics in a conversation," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2021, pp. 2694–2704.
- [8] S. Poria, D. Hazarika, N. Majumder, G. Naik, E. Cambria, and R. Mihalcea, "MELD: A multimodal multi-party dataset for emotion recognition in conversations," in *Proc. Annu. Meeting Assoc. Comput. Linguistics*, 2019, pp. 527–536.
- [9] J. Li et al., "Neutral utterances are also causes: Enhancing conversational causal emotion entailment with social commonsense knowledge," in *Proc. Int. Joint Conf. Artif. Intell.*, 2022, pp. 4209–4215.
- [10] D. Zhang, Z. Yang, F. Meng, X. Chen, and J. Zhou, "Tsam: A two-stream attention model for causal emotion entailment," in *Proc. Int. Conf. Comput. Linguistics*, 2022, pp. 6762–6772.
- [11] P. Ekman, "An argument for basic emotions," *Cogn. & Emotion*, Taylor & Francis, vol. 6, no. 3–4, pp. 169–200, 1992.
- [12] P. P. Liang, Z. Liu, A. B. Zadeh, and L.-P. Morency, "Multimodal language analysis with recurrent multistage fusion," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2018, pp. 150–161.
- [13] Y.-H. H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L.-P. Morency, and R. Salakhutdinov, "Multimodal transformer for unaligned multimodal language sequences," in *Proc. Annu. Meet. Assoc. Comput. Linguist.*, 2019, pp. 6558–6569.
- [14] F. Wang, Z. Ding, R. Xia, Z. Li, and J. Yu, "Multimodal emotion-cause pair extraction in conversations," *IEEE Trans. Affect. Comput.*, vol. 14, no. 3, pp. 1–12, Third Quarter 2023.
- [15] M. Lewis et al., "BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," in *Proc. Annu. Meeting Assoc. Comput. Linguistics*, 2020, pp. 7871–7880.
- [16] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," *Text Summarization Branches Out*, pp. 74–81, 2004.
- [17] Y. Yan, R. Li, S. Wang, F. Zhang, W. Wu, and W. Xu, "ConSERT: A contrastive framework for self-supervised sentence representation transfer," in *Proc. Annu. Meeting Assoc. Comput. Linguistics*, 2021, pp. 5065–5075.
- [18] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: A method for automatic evaluation of machine translation," in *Proc. Annu. Meeting Assoc. Comput. Linguistics*, 2002, pp. 311–318.
- [19] S. Banerjee and A. Lavie, "METEOR: An automatic metric for MT evaluation with improved correlation with human judgments," in *Proc. ACL Workshop Intrinsic Extrinsic Eval. Measures Mach. Transl. Summarization*, 2005, pp. 65–72.
- [20] S. Poria et al., "Recognizing emotion cause in conversations," *Cogn. Computation*, vol. 13, pp. 1317–1332, 2021.
- [21] H. Zhan, T. Sosea, C. Caragea, and J. J. Li, "Why do you feel this way? Summarizing triggers of emotions in social media posts," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2022, pp. 9436–9453.
- [22] D. Ghosal, S. Shen, N. Majumder, R. Mihalcea, and S. Poria, "CICERO: A dataset for contextualized commonsense inference in dialogues," in *Proc. Annu. Meeting Assoc. Comput. Linguistics*, 2022, pp. 5010–5028.
- [23] W. Li, Y. Li, V. Pandeale, M. Ge, L. Zhu, and E. Cambria, "ECPEC: Emotion-cause pair extraction in conversations," *IEEE Trans. Affect. Comput.*, vol. 3, no. 1, pp. 1754–1765, Third Quarter 2023.
- [24] M. Lewis et al., "BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," in *Proc. Annu. Meeting Assoc. Comput. Linguistics*, 2020, pp. 7871–7880.
- [25] A. Vaswani et al., "Attention is all you need," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2017, pp. 6000–6010.
- [26] Y. Liu et al., "Roberta: A robustly optimized bert pretraining approach," 2019, *arXiv: 1907.11692*.
- [27] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "BERTScore: Evaluating text generation with BERT," in *Proc. 8th Int. Conf. Learn. Representations*, Addis Ababa, Ethiopia, Apr. 2020.
- [28] R. Vedantam, C. Lawrence Zitnick, and D. Parikh, "Cider: Consensus-based image description evaluation," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 4566–4575.
- [29] T. Gao, X. Yao, and D. Chen, "SimCSE: Simple contrastive learning of sentence embeddings," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2021, pp. 6894–6910.
- [30] T. Sellam, D. Das, and A. Parikh, "BLEURT: Learning robust metrics for text generation," in *Proc. Annu. Meeting Assoc. Comput. Linguistics*, 2020, pp. 7881–7892.
- [31] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: The munich versatile and fast open-source audio feature extractor," in *Proc. ACM Int. Conf. Multimedia*, 2010, pp. 1459–1462.
- [32] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3D convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 4489–4497.
- [33] J. Hu, Y. Liu, J. Zhao, and Q. Jin, "MMGCN: Multimodal fusion via deep graph convolution network for emotion recognition in conversation," in *Proc. Annu. Meeting Assoc. Comput. Linguistics*, 2021, pp. 5666–5675.
- [34] G. Hu, T.-E. Lin, Y. Zhao, G. Lu, Y. Wu, and Y. Li, "UniMSE: Towards unified multimodal sentiment analysis and emotion recognition," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2022, pp. 7837–7851.
- [35] D. Hu, X. Hou, L. Wei, L. Jiang, and Y. Mo, "MM-DFN: Multimodal dynamic fusion network for emotion recognition in conversations," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2022, pp. 7037–7041.
- [36] L. Ouyang et al., "Training language models to follow instructions with human feedback," 2022, *arXiv:2203.02155*.
- [37] S. Poria, E. Cambria, R. Bajpai, and A. Hussain, "A review of affective computing: From unimodal analysis to multimodal fusion," *Inf. Fusion*, vol. 27, pp. 98–125, 2017.
- [38] D. Hazarika, S. Poria, R. Mihalcea, E. Cambria, and R. Zimmermann, "ICON: Interactive conversational memory network for multimodal emotion detection," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2018, pp. 2594–2604.
- [39] Z. Li, F. Tang, M. Zhao, and Y. Zhu, "EmoCaps: Emotion capsule based model for conversational emotion recognition," in *Proc. Annu. Meet. Assoc. Comput. Linguistics*, 2022, pp. 1610–1618.
- [40] J. Li, X. Wang, G. Lv, and Z. Zeng, "GA2MIF: Graph and attention based two-stage multi-source information fusion for conversational emotion detection," *IEEE Trans. Affect. Comput.*, vol. 15, no. 1, pp. 130–143, First Quarter 2024.
- [41] S. Y. M. Lee, Y. Chen, and C.-R. Huang, "A text-driven rule-based system for emotion cause detection," in *Proc. Workshop Comput. Approaches Anal. Gener. Emotion Text*, 2010, pp. 45–53.
- [42] R. Xia and Z. Ding, "Emotion-cause pair extraction: A new task to emotion analysis in texts," in *Proc. Annu. Meeting Assoc. Comput. Linguistics*, 2019, pp. 1003–1012.
- [43] L. Gui, R. Xu, Q. Lu, D. Wu, and Y. Zhou, "Emotion cause extraction, a challenging task with corpus construction," in *Proc. 5th Nat. Conf. Social Media Process.*, 2016, pp. 98–109.
- [44] W. Zhao, Y. Zhao, Z. Li, and B. Qin, "Knowledge-bridged causal interaction network for causal emotion entailment," in *Proc. AAAI Conf. Artif. Intell.*, 2023, pp. 14020–14028.
- [45] J. Li, X. Wang, G. Lv, and Z. Zeng, "GraphCFC: A directed graph based cross-modal feature complementation approach for multimodal conversational emotion recognition," 2022, *arXiv:2207.12261*.
- [46] M. Riyadh and M. O. Shafiq, "Towards emotion cause generation in natural language processing using deep learning," in *Proc. IEEE Int. Conf. Mach. Learn. Appl.*, 2022, pp. 140–147.
- [47] V. Chudasama, P. Kar, A. Gudmalwar, N. Shah, P. Wasnik, and N. Onoe, "M2fnet: Multi-modal fusion network for emotion recognition in conversation," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recogn. Workshops*, 2022, pp. 4651–4660.
- [48] X. Zhao, Y. Chen, S. Liu, X. Zang, Y. Xiang, and B. Tang, "Tmmda: A new token mixup multimodal data augmentation for multimodal sentiment analysis," in *Proc. ACM Web Conf.*, 2023, pp. 1714–1722.
- [49] S. Ge, Z. Jiang, Z. Cheng, C. Wang, Y. Yin, and Q. Gu, "Learning robust multi-modal representation for multi-label emotion recognition via adversarial masking and perturbation," in *Proc. ACM Web Conf.*, 2023, pp. 1510–1518.
- [50] J. Cohen, "A coefficient of agreement for nominal scales," *Educ. Psychol. Meas.*, vol. 20, pp. 37–46, 1960.
- [51] M. L. McHugh, "Interrater reliability: The kappa statistic," *Biochemia Medica*, vol. 22, pp. 276–282, 2012.

- [52] G. Tu, B. Liang, D. Jiang, and R. Xu, "Sentiment- emotion- and context-guided knowledge selection framework for emotion recognition in conversations," *IEEE Trans. Affect. Comput.*, vol. 14, no. 3, pp. 1803–1816, Third Quarter 2023.
- [53] S. Tang, C. Wang, K. Xu, Z. Huang, M. Xu, and Y. Peng, "An emotion evolution network for emotion recognition in conversation," in *Proc. Int. Conf. Tools Artif. Intell.*, 2022, pp. 1231–1238.
- [54] Y. Kang and Y.-S. Cho, "Directed acyclic graphs with prototypical networks for few-shot emotion recognition in conversation," *IEEE Access*, pp. 117633–117642, 2023.
- [55] Q. Guo, J. Yu, Y. Zhang, H. Jiang, W. Liu, and J. Yin, "Discovery of emotion implicit causes in products based on commonsense reasoning," in *Proc. Int. Conf. Adv. Data Mining Appl.*, 2023, pp. 277–292.
- [56] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "HuBER: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 29, pp. 3451–3460, 2021.
- [57] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.
- [58] S. Patel, D. Shukla, and A. Modi, "Litk at semeval-2024 task 10: Who is the speaker? Improving emotion recognition and flip reasoning in conversations via speaker embeddings," 2024, *arXiv:2404.04525*.
- [59] G. Singh, D. Brahma, P. Rai, and A. Modi, "Text-based fine-grained emotion prediction," *IEEE Trans. Affect. Comput.*, vol. 15, no. 2, pp. 405–416, Second Quarter 2024.
- [60] K. Bansal, H. Agarwal, A. Joshi, and A. Modi, "Shapes of emotions: Multimodal emotion recognition in conversations via emotion shifts," in *Proc. Int. Conf. Comput. Linguistics*, 2022, pp. 44–56.
- [61] A. Bhat and A. Modi, "Multi-task learning framework for extracting emotion cause span and entailment in conversations," in *Proc. Workshop Transfer Learn. Natural Lang. Process.*, 2023, pp. 33–51.
- [62] F. Wang, J. Yu, and R. Xia, "Generative emotion cause triplet extraction in conversations with commonsense knowledge," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2023, pp. 3952–3963.
- [63] F. Wang, H. Ma, J. Yu, R. Xia, and E. Cambria, "Semeval-2024 task 3: Multimodal emotion cause analysis in conversations," 2024, *arXiv:2405.13049*.
- [64] F. Wang, H. Ma, X. Shen, J. Yu, and R. Xia, "Observe before generate: Emotion-cause aware video caption for multimodal emotion cause generation in conversations," in *Proc. ACM Multimedia*, 2024.
- [65] P. Colombo, W. Witon, A. Modi, J. Kennedy, and M. Kapadia, "Affect-driven dialog generation," 2019, *arXiv: 1904.02793*.
- [66] I. Singh, A. Barkati, T. Goswamy, and A. Modi, "Adapting a language model for controlled affective text generation," 2020, *arXiv: 2011.04000*.



Heqing Ma received the BS degree in intelligent science and technology from the Nanjing University of Science and Technology, Nanjing, China, in 2022. She is currently working toward the MS degree with the School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing, China. Her research interests include natural language processing and effective computing.



Jianfei Yu received the BS and MS degrees from the Nanjing University of Science and Technology, Nanjing, China, in 2012 and 2015, respectively, and the PhD degree from Singapore Management University, Singapore, in 2018. He is currently an associate professor with the School of Computer Science and Engineering, Nanjing University of Science and Technology, China. His research interests include natural language processing, machine learning, and data mining.



Fanfan Wang received the BS degree in automation from the Nanjing University of Science and Technology, Nanjing, China, in 2019. She is currently working toward the PhD degree in computer science and technology with the School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing, China. Her research interests include natural language processing and effective computing.



Hanyu Cao received the BS degree in robotics engineering from the Nanjing University of Information Science and Technology, Nanjing, China, in 2023. She is currently working toward the MS degree with the School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing, China. Her research interests include natural language processing and effective computing.



Rui Xia received the PhD degree from the Institute of Automation, Chinese Academy of Sciences, in 2011. He is currently a professor with the School of Computer Science and Engineering, Nanjing University of Science and Technology, China. His research interests include natural language processing, data mining, and affective computing. He has published more than 50 papers in top journals and conferences. His work on emotion-cause pair extraction has received the ACL2019 Outstanding Paper Award.