SemEval-2024 Task 3: Multimodal Emotion Cause Analysis in Conversations

Fanfan Wang[1], **Heqing Ma**[1], **Jianfei Yu**[1*], **Rui Xia**[1*], **Erik Cambria**[2]
[1] School of Computer Science and Engineering,
Nanjing University of Science and Technology, China
[2] Nanyang Technological University, Singapore
{ffwang, hqma, jfyu, rxia}@njust.edu.cn, cambria@ntu.edu.sg

Abstract

The ability to understand emotions is an essential component of human-like artificial intelligence, as emotions greatly influence human cognition, decision making, and social interactions. In addition to emotion recognition in conversations, the task of identifying the potential causes behind an individual's emotional state in conversations, is of great importance in many application scenarios. We organize SemEval-2024 Task 3, named Multimodal Emotion Cause Analysis in Conversations, which aims at extracting all pairs of emotions and their corresponding causes from conversations. Under different modality settings, it consists of two subtasks: Textual Emotion-Cause Pair Extraction in Conversations (TECPE) and Multimodal Emotion-Cause Pair Extraction in Conversations (MECPE). The shared task has attracted 143 registrations and 216 successful submissions. In this paper, we introduce the task, dataset and evaluation settings, summarize the systems of the top teams, and discuss the findings of the participants.

1 Introduction

Understanding emotions is crucial to achieve human-like artificial intelligence, as emotions are intrinsic to humans and significantly influence our cognition, decision-making, and social interactions. Conversation is an important form of human communication and contains a large number of emotions. Furthermore, given that conversation in its natural form is multimodal, many studies have explored multimodal emotion recognition in conversations (ERC), using language, audio and vision modalities (Poria et al., 2019b; Mittal et al., 2020; Lian et al., 2021; Zhao et al., 2022; Zheng et al., 2023).

However, emotion recognition alone is not sufficient to fully understand the intricacies of human emotions. Emotion cause analysis (ECA), the process of identifying the potential causes behind an individual's emotion state, has broad application scenarios such as human-computer interaction, commerce customer service, empathetic conversational agents, and automatic psychotherapy. For example, conversational agents equipped with emotion cause analysis can better understand the user's emotional state, offer empathetic responses, and provide more personalized services. By identifying the cause of the emotional state of a patient, a psychotherapy system can provide more accurate and customized treatments. ECA has gained increasing attention both in academic and practical fields (Ding et al., 2019; Xia et al., 2019; Xia and Ding, 2019; Ding et al., 2020a,b; Poria et al., 2021; Li et al., 2022; An et al., 2023; Wang et al., 2023b). However, to our knowledge, there has not been any evaluation competition conducted specifically for emotion cause analysis in conversations.

To promote research in this direction, we organize a shared task in SemEval-2024, named Multimodal Emotion Cause Analysis in Conversations. Our task consists of two subtasks: Subtask 1 (Textual Emotion-Cause Pair Extraction in Conversations, TECPE) focuses on extracting emotion and textual cause spans solely based on text; Subtask 2 (Multimodal Emotion-Cause Pair Extraction in Conversations, MECPE) involves extracting emotion-cause pairs at the utterance level considering three modalities.

For this shared task, we provide a multimodal emotion cause dataset ECF 2.0 sourced from the sitcom *Friends*. This dataset contains 1,715 conversations and 16,720 utterances, where 12,256 emotion-cause pairs are annotated at the utterance level, covering three modalities (language, audio, and vision). Specifically, in our preliminary work (Wang et al., 2023a), we have constructed a benchmark dataset, Emotion-Cause-in-Friends (ECF 1.0), which contains 1,374 conver-

* Corresponding authors.

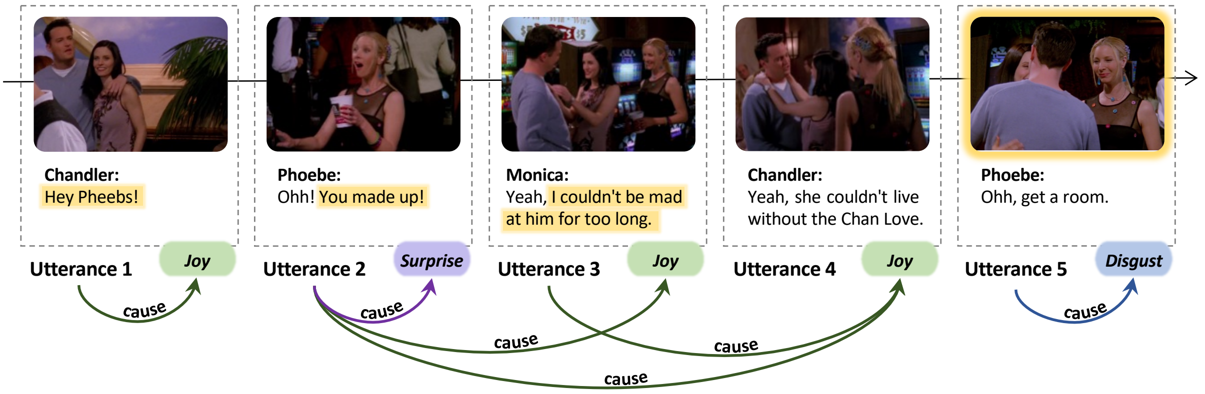


Figure 1: An example of our task and annotated dataset. Each arc points from the cause utterance to the emotion it triggers. The textual cause spans and the visual cause evidence are highlighted in yellow. Background: Chandler and his girlfriend Monica walked into the casino (they had a quarrel earlier but made up soon) and then started a conversation with Phoebe.

sations and 13,619 utterances. On this basis, we have furthermore annotated an extended test set as the evaluation data and provided the span-level annotations of emotion causes within the textual modality.

Our task has attracted 143 registrations and a total of 216 successful submissions during the 16-day evaluation phase. Participants tended to decompose our task into emotion recognition and cause prediction, proposing numerous well-designed pipeline systems. Moreover, many teams applied advanced Large Language Models (LLMs) for emotion cause analysis and achieved promising results. After the evaluation, 18 teams finally submitted system description papers.

2 Task

We clarify the definitions of emotion and cause before introducing the task and dataset. **Emotion** is a psychological state associated with thought, feeling, and behavioral response (Ekman and Davidson, 1994). In computer science, emotions are often described as discrete emotion categories, such as Ekman's six basic emotions, including *Anger*, *Disgust*, *Fear*, *Joy*, *Sadness* and *Surprise* (Ekman, 1971). In conversations, emotions were usually annotated at the utterance level (Li et al., 2017; Hsu et al., 2018; Poria et al., 2019a). **Cause** refers to the objective event or subjective argument that triggers the corresponding emotion (Lee et al., 2010; Russo et al., 2011).

The goal of our shared task, named Multimodal Emotion Cause Analysis in Conversations, is to extract potential pairs of emotions and their corresponding causes from a given conversation. Figure 1 illustrates a typical multimodal conversation scenario, which involves multiple emotions and their corresponding causes. Under different modality settings, we define the following two subtasks:

Subtask 1: Textual Emotion-Cause Pair Extraction in Conversations (TECPE). Extracting all emotion-cause pairs from the given conversation solely based on text, where each pair contains an emotion utterance along with its emotion category and the textual cause span, e.g., (*U3_Joy*, *U2_*"*You made up!*") in Figure 1.

Subtask 2: Multimodal Emotion-Cause Pair Extraction in Conversations (MECPE). It should be noted that sometimes the cause cannot be reflected only in text. As shown in Figure 1, the cause for Phoebe's *Disgust* in *U5* is that Monica and Chandler were kissing in front of her, which is reflected in the visual modality of *U5*. Therefore, we accordingly define this multimodal subtask to extract all emotion-cause pairs in consideration of three modalities (language, audio, and vision). In this subtask, the cause is defined at the utterance level, and each pair contains an emotion utterance along with its emotion category and a cause utterance, e.g., (*U5_Disgust*, *U5*).

3 Dataset

3.1 Data Source

Sitcoms come with real-world-inspired interhuman interactions and usually contain more emotions than other TV series or movies. Based on the famous American sitcom *Friends*, Poria et al. (2019a) constructed the multimodal conversational dataset MELD by extracting audiovisual clips corresponding to the scripts of the source episodes and annotating each utterance with one of six basic emotions (*Anger*, *Disgust*, *Fear*, *Joy*, *Sadness* and *Surprise*) or *Neutral*. MELD has recently become a widely used benchmark for ERC.

In our preliminary work (Wang et al., 2023a), we chose MELD as the data source and further annotated the causes given emotion annotations, thereby constructing the ECF 1.0 dataset. For this SemEval competition, we release the entire ECF 1.0 dataset as a training set and additionally create a test set as evaluation data, which is also sourced from *Friends*.

3.2 Data Collection

To construct the extended test set, we first crawl the subtitle files of all the episodes of *Friends*, which contains the utterance text and the corresponding timestamps. The subtitles are then separated by scene (scene descriptions are written in square brackets in the subtitle files), and each scene in every episode is viewed as a conversation. If the length of a conversation exceeds 40 utterances, we further divide it into several conversations of random lengths. Conversations included in the ECF 1.0 are removed. Next, we divide the collected conversations into several parts according to their lengths, with each part falling within the length ranges [1, 5], [6, 10], [11, 15], [16, 20], [21, 25], and [26, 35], respectively. Finally, we randomly sample conversations from each part according to the distribution probability of conversation lengths in ECF 1.0, and a total of 400 conversations are sampled for annotation.

3.3 Data Annotation

We employ three graduate students involved in the annotation of the ECF 1.0 dataset to annotate the extended test set. Given a multimodal conversation, they first need to annotate the speaker and emotion category for each utterance, and then further annotate the utterances containing corresponding causes for each non-neutral emotion. If the causes are explicitly expressed in the text, they should also mark the textual cause spans. After annotation, we determine the emotion categories and cause utterances by majority voting, and take the largest boundary (i.e., the union of the spans) as the gold annotation of the textual cause span. If disagreements arise, another expert is invited for the final decision.

Dataset	Modality	Scene	# Ins
Emotion-Stimulus (Ghazi et al., 2015)	T	–	2,414 s
ECE Corpus (Gui et al., 2016)	T	News	2,105 d
NTCIR-13-ECA (Gao et al., 2017)	T	Fiction	2,403 d
Weibo-Emotion (Cheng et al., 2017)	T	Blog	7,000 p
REMAN (Kim and Klinger, 2018)	T	Fiction	1,720 d
GoodNewsEveryone (Bostan et al., 2020)	T	News	5,000 s
RECCON-IE (Poria et al., 2021)	T	Conv	665 u
RECCON-DD (Poria et al., 2021)	T	Conv	11,104 u
ConvECPE (Li et al., 2022)	T,A,V	Conv	7,433 u
ECF 1.0 (Wang et al., 2023a)	T,A,V	Conv	13,619 u
ECF 2.0	T,A,V	Conv	16,720 u

Table 1: Comparison of existing ECA datasets. T, A, and V refer to text, audio, and video. Blog and Conv represent microblog and conversation, and s, d, p and u denote sentence, document, post and utterance.

Items	ECF 1.0	Extended Test	ECF 2.0
Conversations	1,374	341	1,715
Utterances	13,619	3,101	16,720
Emotion (utterances)	7,690	1,821	9,511
Subtask 1 (TECPE)			
Emotion (utterances) with causes	6,761	1,626	8,387
Emotion-cause (span) pairs	9,284	2,256	11,540
Subtask 2 (MECPE)			
Emotion (utterances) with causes	7,081	1,746	8,827
Emotion-cause (utterance) pairs	9,794	2,462	12,256

Table 2: Statistics of our dataset.

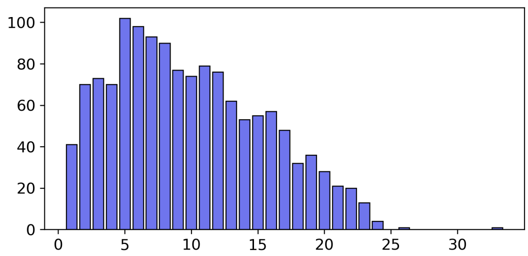
Annotation Cost. The average duration of each conversation in our dataset is 31.6 seconds and it takes about 10 minutes to annotate a conversation. Each annotator would be paid CNY 300 when finishing every 50 conversations, which leads to the basic salary of CNY 36 (USD 5.2) per hour, which is higher than the current average salary in Jiangsu Province, China.

Data Post-processing. We conduct the following post-processing and cleaning of the data:

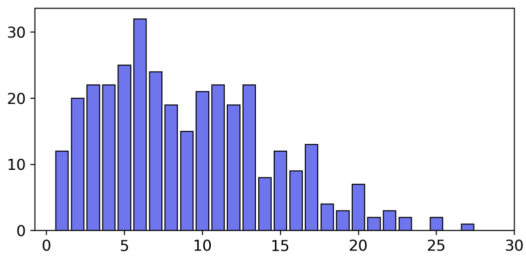
- Correct the utterance text that does not match what the speaker said in the video;
- Correct the timestamps that are not aligned with utterance text;
- Separate the utterance whose segment of timestamps covers two speakers' utterances and modify their timestamps;
- Separate the conversation which spans scenes;
- Discard conversations if there is significant disagreement in annotations and the expert also finds it difficult to determine.

After these steps, we store the text data in JSON files separately for each subtask. For Subtask 2, we use the FFmpeg[1] tool to extract video clips of

[1]https://www.ffmpeg.org

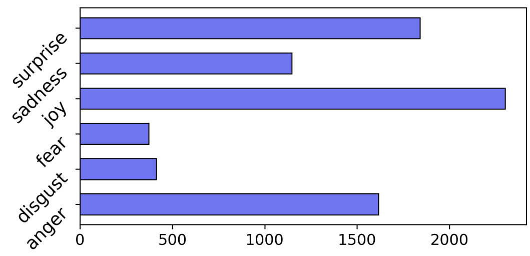


(a) ECF 1.0

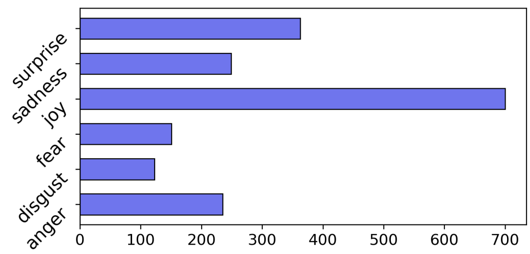


(b) Extended Test set for SemEval-2024

Figure 2: The distribution of conversation lengths. The horizontal axis represents the number of utterances, and the vertical axis represents the number of conversations.



(a) ECF 1.0



(b) Extended Test set for SemEval-2024

Figure 3: The distribution of emotions. The horizontal axis represents the number of utterances, and the vertical axis represents emotion categories.

each utterance from the source episodes based on the start and end timestamps.

3.4 Dataset Statistic

In our preliminary work (Wang et al., 2023a), we have already constructed the ECF 1.0 dataset that contains 1,374 conversations and 13,619 utterances. Furthermore, we have annotated an extended test set specifically for this SemEval evaluation, which together with ECF 1.0 constitutes the **ECF 2.0** dataset[2] that contains 1,715 conversations and 16,720 utterances.

In Table 1, we compare our dataset with the related datasets for ECA, in terms of modality, scene, and size. It is evident that ECF 2.0 is currently the largest available emotion cause dataset.

Table 2 presents the detailed statistics of our dataset for the two subtasks. It can be seen that, in the entire ECF 2.0 dataset, 56.88% of the utterances are labeled with one of the six basic emotions, 92.81% of the emotion utterances have corresponding cause utterances, and 88.18% of the emotion utterances are annotated with textual cause spans.

In addition, as shown in Figure 2 and Figure 3, the newly annotated test set is basically consistent with the original ECF 1.0 dataset in terms of conversation length and emotion distribution.

4 Evaluation

Our SemEval task runs on CodaLab[3]. We released the training data in September 2023, and notified participants to commence model development. The evaluation phase began on January 16, 2024, and ended on January 31, 2024. We mixed the extended test set (consisting of 341 conversations with emotion and cause annotations; the labels are not publicly available) with some noise data (containing 324 conversations, not intended for evaluation) and released them together. Each team is allowed to submit their results up to three times a day.

4.1 Evaluation Metrics

We evaluate the emotion-cause pairs of each emotion category with F_1 scores separately and further calculate a weighted average of F_1 scores across the six emotion categories, denoted as "**w-avg. F_1**". Specifically, for Subtask 1, which involves the textual cause span, we adopt two strategies to determine whether the span is extracted correctly:

- *Strict Match*: A predicted span is regarded as correct if it's the same as one of the annotated spans;

[2] Our dataset is available on Google Drive.

[3] https://codalab.lisn.upsaclay.fr/competitions/16141

Rank	User Name	Team Name	w-avg. S. F_1	w-avg. P. F_1	Main Technologies
1	Mercurialzs	Samsung Research China-Beijing†	0.2300	0.3223	LLaMA2, SpanBERT
2	sachertort	petkaz†	0.1035	0.2640	GPT 3.5, BERT
3	sharadC	UIC NLP GRADS†	0.1839	0.2442	RoBERTa, SpanBERT
4	nicolay-r	nicolay-r†	0.1279	0.2432	Flan-T5
5	Mahshid	AIMA†	0.0218	0.2102	EmoBERTa, DeBERTa
6	jimar	UWBA†	0.0639	0.2084	RoBERTa, BERT
7	Choloe_guo	UIR-ISC†	0.1518	0.1963	BERT, SpanBERT
8	aranjan25	–	0.1431	0.1930	–
9	anaezquerro	LyS†	0.0677	0.1823	BERT
10	wrafal	PWEITINLP†	0.0449	0.0723	GPT-3, SpanBERT
11	ericcui	(͡° ͜ʖ ͡°)-GPT	0.0033	0.0339	–
12	conner	–	0.0000	0.0063	–
13	hpiotr6	–	0.0000	0.0046	–
14	deliagrigorita	–	0.0005	0.0024	–
15	jpcf12	VerbaNexAI Lab†	0.0000	0.0000	Logistic Regression, SpaCy

Table 3: The leaderboard for Subtask 1 (TECPE). "†" indicates that the team has submitted a system description paper to SemEval-2024.

Rank	User Name	Team Name	w-avg. F_1	Modality	Main Technologies
1	Mercurialzs	Samsung Research China-Beijing†	0.3774	T,A,V	LLaMA2, RoBERTa, LLaVA
2	ZhanG_XD	NUS-Emo†	0.3460	T,V	ChatGLM3
3	SZTU-MIPS	SZTU-MIPS†	0.3435	T,A,V	MiniGPT-v2
4	arefa	JMI†	0.2758	T,V	GPT-4V, GPT-3.5
5	Mahshid	AIMA†	0.2584	T	EmoBERTa
6	jimar	UWBA†	0.2506	T,A,V	RoBERTa, BERT
7	julia-bel	DeepPavlov†	0.2057	T,A,V	Video-LLaMA
8	akshettrj	LastResort†	0.1836	T	BiLSTM, CRF
9	oliver_wang	QFNU_CS†	0.1786	T,A,V	BERT
10	MSurfer20	–	0.1708	–	–
11	ayushg2000	–	0.1635	–	–
12	Hidetsune	Hidetsune†	0.1288	T	SpaCy, BERT
13	DuyguA	D-NLP	0.0521	–	–
14	bbgame605065444	NCL†	0.0146	T,A,V	MLP
15	joshuashunk	–	0.0008	–	–

Table 4: The leaderboard for Subtask 2 (MECPE). "†" indicates that the team has submitted a system description paper to SemEval-2024.

- *Proportional Match*: Calculate the overlap proportion of the predicted span and the annotated one.

The evaluation metrics for the two strategies are "w-avg. S. F_1" and "w-avg. P. F_1", respectively. Taking into account the complexity of Subtask 1, we choose "**w-avg. P. F_1**" as the main metric[4] for the ranking.

4.2 Baselines

As mentioned in our previous work (Wang et al., 2023a), for Subtask 2 we also employed the BiLSTM-based ECPE-2steps model as our baseline system. Specifically, we maintain the validation set of the ECF 1.0 datset unchanged and merge the test set into the training set to train the model. The evaluation of the model predictions on the extended test set achieves a weighted average F_1 of **0.1926**.

For Subtask 1, based on the same model, we just convert the cause extraction module in Step 1 from the cause utterance prediction to the prediction of the start index and end index within the utterance, then simply match the indexes as candidate cause spans, followed by emotion-cause pairing and filtering in Step 2. The evaluation result for the weighted average proportional F_1 on the extended test set is **0.1801**.

4.3 Participating Systems and Results

Our competition was created on Codalab in November 2023, and has attracted 143 registrations and a total of 216 submissions. After the evaluation, 18 teams have submitted system de-

[4]Specific calculation details can be found on GitHub.

scription papers.

Team *Samsung Research China-Beijing* (Zhang et al., 2024) won first place in both subtasks, holding a significant lead over the second-place team. Teams *petkaz* (Kazakov et al., 2024) and *UIC NLP GRADS* (Chandakacherla et al., 2024) respectively captured the second and third places in Subtask 1. Teams *NUS-Emo* (Luo et al., 2024) and *SZTU-MIPS* (Cheng et al., 2024) attained second and third positions in Subtask 2. The official leaderboards for Subtask 1 and Subtask 2 are shown in Table 3 and Table 4, respectively.

4.3.1 System Overview

Almost all systems have implemented our task through a two-step framework, first performing the ERC task and then predicting the causes based on emotions. In the following, we briefly introduce the systems from the top teams and some other notable approaches.

Team *Samsung Research China-Beijing* (Zhang et al., 2024) achieved first place in both subtasks with a pipeline framework. They fine-tuned the LLaMA2-based InstructERC (Lei et al., 2023) to extract the emotion category of each utterance in a conversation. For further data augmentation, they added three additional auxiliary tasks based on the original training data strategy of InstructERC. Then, the MuTEC (Bhat and Modi, 2023) and TSAM (Zhang et al., 2022) models are used, respectively, to extract cause spans for Subtask 1 and cause utterances for Subtask 2. They also obtained different multimodal representations through openSMILE (Eyben et al., 2010), LLaVA (Liu et al., 2024), and a self-designed face module to explore the integration of audio-visual information. It should be noted that they used various models for ensemble learning to determine the final prediction.

Team *petkaz* (Kazakov et al., 2024) ranked second in Subtask 1. They fine-tuned GPT 3.5 (Ouyang et al., 2022) for emotion classification and then used a BiLSTM-based neural network to detect cause utterances. The cause extractor model is initialized with BERT (Devlin et al., 2019), followed by three BiLSTM layers. They treat the entire cause utterance as a cause span.

Team *NUS-Emo* (Luo et al., 2024) achieved the second highest score in Subtask 2. First, they conducted zero-shot testing experiments to evaluate multiple LLMs, including OPT-IML3 (Iyer et al., 2022), Instruct-GPT4 (Peng et al., 2023), Flan-T5 (Chung et al., 2022), and ChatGLM (Du et al., 2022). ChatGLM3-6B is ultimately selected as its backbone model based on its superior performance. They designed an emotion-cause-aware instruction-tuning mechanism to update the LLM and incorporated the visual representation from the ImageBind (Girdhar et al., 2023) encoder.

Team *UIC NLP GRADS* (Chandakacherla et al., 2024) achieved the third place in Subtask 1, and their system performed well in the strict metric, ranking second. They fine-tuned RoBERTa (Liu et al., 2019) for emotion classification, and then further fine-tuned a SpanBERT (Joshi et al., 2019) model that had been fine-tuned in SQuAD 2.0 (Rajpurkar et al., 2018), to predict cause spans in QA format.

Team *SZTU-MIPS* (Cheng et al., 2024) ranked third in Subtask 2. They integrated text, audio, and image modalities for emotion recognition and adopted the MiniGPTv2 model (Chen et al., 2023) for multimodal cause extraction. Specifically, textual features are obtained from InstructERC, while acoustic features are extracted using HuBERT (Hsu et al., 2021). For visual modality, faces are first extracted using OpenFace (Baltrusaitis et al., 2016) from video frames, followed by extraction of facial features using expMAE (Cheng et al., 2023).

Team *nicolay-r* (Rusnachenko and Liang, 2024) finetuned Flan-T5 by designing the chain of thoughts for emotion causes based on the Three-Hop Reasoning (THOR) framework (Fei et al., 2023), to predict the emotion of the current utterance and the emotion caused by the current utterance towards the target utterance. Their reasoning revision methodology and rule-based span correction technique bring further improvements.

Team *JMI* (. et al., 2024) implemented two different approaches. In their best system, they used in-context learning using GPT 3.5 for emotion prediction and cause prediction, respectively. Conversation-level video descriptions were extracted via GPT-4V (Yang et al., 2023) to provide more context to GPT 3.5. In addition, they also fine-tuned two separate Llama2 (Touvron et al., 2023) models to recognize emotions and extract causes.

Team *AIMA* (Ghahramani Kure et al., 2024) fine-tuned EmoBERTa (Kim and Vossen, 2021) for emotion classification and then obtained the emotion-cause pairs via a Transformer-based encoder. After finding the pairs, they further fine-tuned the DeBERTa (He et al., 2021) that had been fine-tuned on SQuAD 2.0 to extract the cause spans for Subtask 1.

Team *UWBA* (Baloun et al., 2024) fused the features of three modalities at the utterance level and then used them for emotion classification and pair prediction. It is interesting that they summarized five span categories (*Whole Utterance*, *First part*, *Last part*, *Middle part*, *Other*) through observations of training data, and then trained a classifier to further predict textual cause spans in cause utterance.

Furthermore, Team *DeepPavlov* (Belikova and Kosenko, 2024) investigated the performance of Video-LLaMA (Zhang et al., 2023) in several modes and found that model fine-tuning yields notable improvements in emotion and cause classification. Team *PWEITINLP* (Levchenko et al., 2024) utilized GPT-3 for emotion classification. Some other Teams, including *UIR-ISC* (Guo et al., 2024), *LyS* (Ezquerro and Vilares, 2024), *QFNU_CS* (Wang et al., 2024) and *Hidetsune* (Takahashi, 2024), all employed BERT-based models to address our task, among which *LyS* proposed an end-to-end model comprising a BERT encoder and a graph-based decoder to identify emotion cause relations. Team *LastResort* (Mathur et al., 2024) tackled our task as sequence labeling problems and used BiLSTM followed by a CRF layer to solve it. Team *NCL* (Li et al., 2024) solely utilized pre-trained models to extract features from three modalities. Team *VerbaNexAI Lab* (Pacheco et al., 2024) demonstrated the inadequacy of machine learning techniques alone for emotion cause analysis.

4.3.2 Discussion

Our task, Multimodal Emotion Cause Analysis in Conversations, involves informal real-life conversations and complex audio-visual scenes. Additionally, emotions exhibit strong subjectivity, and we have observed that even humans sometimes struggle to accurately identify emotions and their causes. This complexity underscores the intricate nature of human emotions and the nuanced contexts in which they occur, posing a substantial challenge for data annotation and subsequent model development.

Dataset Bias. Emotion category imbalance is an inherent problem in the ERC task (Li et al., 2017; Hsu et al., 2018; Poria et al., 2019a), aligning with real-world phenomena where people tend to express positive emotions like *joy* more frequently in their daily communications, while expressions of *disgust* and *fear* are less common. Our dataset is sourced from TV series that closely resemble the real world, naturally also exhibiting an imbalance in emotions, as illustrated in Figure 3. However, such an imbalance may adversely affect a model's ability to learn and generalize across different emotions, potentially leading to biases towards frequently expressed emotions (Kazakov et al., 2024; Chandakacherla et al., 2024). Moreover, emotion cause datasets often have a noticeable pattern in the location of causes and emotions. Some systems rely on this position bias, either by using a fixed window size or by direct post-processing to add the emotion utterance as the cause (Rusnachenko and Liang, 2024; . et al., 2024), which overlooks the effective semantic connections between distant contexts and may lead to poor generalization capabilities for unseen data where the cause is not in proximity to the emotion. In the future, LLMs can be leveraged to assist with annotation to expand the diversity of datasets available for fine-tuning, which encompass a wider range of emotional expressions and cultural backgrounds. This can mitigate existing dataset biases and enhance the model's applicability and generalizability across various scenarios.

Utilization of LLMs. Recently, LLMs have exhibited remarkable capabilities in a wide range of tasks and are rapidly advancing the field of natural language processing. Therefore, LLMs are allowed to be used in our competition. It is evident that about a third of the teams have used LLMs for emotion cause analysis, and most of them are ranked at the top. However, some participants have observed that LLMs perform poorly in zero-shot and few-shot settings on emotion and cause recognition tasks (Kazakov et al., 2024; . et al., 2024; Belikova and Kosenko, 2024), indicating a crucial need for task-specific fine-tuning. Furthermore, prompt engineering is essential, as LLMs often produce hallucinations or unstructured outputs. Due to resource and cost constraints, most researchers cannot take full advantage of the strongest capabilities of LLM. Future research is encouraged to explore ways to enhance lightweight models or to bridge the gap between pre-training and downstream tasks, thereby augmenting LLMs' ability to understand emotions.

Potential of Multimodal Information. Multimodal information is important for discovering both emotions and their causes in conversations. In our daily communications, we depend not only on the speaker's voice intonation and facial expressions to perceive his emotions, but also on some auditory and visual scenes to speculate the potential causes that trigger the emotions of speakers beyond text. However, some participants found that the introduction of audio or visual modalities results in minimal improvements or even a decrease in system performance (Zhang et al., 2024; Cheng et al., 2024; Baloun et al., 2024). This issue arises partly due to the characteristics of our dataset, which involves a large number of complex visual scenes but few visual cause clues, leading to the introduction of noise. Another limiting factor might be that multimodal feature extraction methods are not advanced enough or fusion strategies are not effective enough. The challenges that require further exploration include the effective interaction and fusion of multimodal information, as well as the perception, understanding, and utilization of audiovisual scenes. Furthermore, there is a demand for more high-quality data sets on multimodal emotion cause analysis to support research in this area.

5 Conclusions

In this paper, we describe the SemEval-2024 Task 3 named Multimodal Emotion Cause Analysis in Conversations, which aims to extract all potential pairs of emotions and their corresponding causes from a conversation. The shared task has attracted 143 registrations and 216 successful submissions. We provide detailed descriptions of task definition and data annotation, summarize participating systems, and discuss their findings.

As an important direction of affective computing, multimodal emotion cause analysis in conversation plays an important role in many real-world applications. We hope that our research and resources can contribute towards the design of future systems in this direction.

6 Ethics Statement

Our **ECF 2.0** dataset is annotated on the basis of the MELD dataset [5] which is licensed under the GNU General Public License v3.0 and is used only for scientific research. We do not share personal information and do not release sensitive content that can be harmful to any individual or community. Conducting multimodal emotion cause analysis will help us better understand emotions in human conversations, build human-machine dialogue systems, and contribute to society and human well-being.

[5] https://github.com/declare-lab/MELD

Acknowledgements

We express our sincere gratitude to the annotators who contributed to constructing the dataset, laying a solid foundation for our research. We also extend our heartfelt gratitude to all the participants who participated in our competition, especially the teams who submitted system description papers and completed the reviewing tasks assigned to them. Furthermore, we thank the anonymous reviewers for their invaluable feedback and insightful comments.

References

Arefa ., Mohammed Abbas Ansari, Chandni Saxena, and Tanvir Ahmad. 2024. JMI at SemEval 2024 task 3: Two-step approach for multimodal ECAC using in-context learning with GPT and instruction-tuned llama models. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 1561–1576, Mexico City, Mexico. Association for Computational Linguistics.

Jiaming An, Zixiang Ding, Ke Li, and Rui Xia. 2023. Global-view and speaker-aware emotion cause extraction in conversations. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

Josef Baloun, Jiri Martinek, Ladislav Lenc, Pavel Kral, Matěj Zeman, and Lukáš Vlček. 2024. UWBA at SemEval-2024 task 3: Dialogue representation and multimodal fusion for emotion cause analysis. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 316–325, Mexico City, Mexico. Association for Computational Linguistics.

Tadas Baltrusaitis, Peter Robinson, and Louis-Philippe Morency. 2016. Openface: An open source facial behavior analysis toolkit. *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1–10.

Julia Belikova and Dmitrii Kosenko. 2024. DeepPavlov at SemEval-2024 task 3: Multimodal large language models in emotion reasoning. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 1747–1757, Mexico City, Mexico. Association for Computational Linguistics.

Ashwani Bhat and Ashutosh Modi. 2023. Multi-task learning framework for extracting emotion cause span and entailment in conversations. In *Transfer Learning for Natural Language Processing Workshop*, pages 33–51. PMLR.

Laura Ana Maria Bostan, Evgeny Kim, and Roman Klinger. 2020. Goodnewseveryone: A corpus of news headlines annotated with emotions, semantic roles, and reader perception. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 1554–1566.

Sharad Chandakacherla, Vaibhav Bhargava, and Natalie Parde. 2024. UIC NLP GRADS at SemEval-2024 task 3: Two-step disjoint modeling for emotion-cause pair extraction. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 1373–1379, Mexico City, Mexico. Association for Computational Linguistics.

Jun Chen, Deyao Zhu, Xiaoqian Shen, Xiang Li, Zechun Liu, Pengchuan Zhang, Raghuraman Krishnamoorthi, Vikas Chandra, Yunyang Xiong, and Mohamed Elhoseiny. 2023. Minigpt-v2: large language model as a unified interface for vision-language multi-task learning. *ArXiv*, abs/2310.09478.

Xiyao Cheng, Ying Chen, Bixiao Cheng, Shoushan Li, and Guodong Zhou. 2017. An emotion cause corpus for chinese microblogs with multiple-user structures. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, 17(1):1–19.

Zebang Cheng, Yuxiang Lin, Zhaoru Chen, Xiang Li, Shuyi Mao, Fan Zhang, Daijun Ding, Bowen Zhang, and Xiaojiang Peng. 2023. Semi-supervised multimodal emotion recognition with expression mae. *Proceedings of the 31st ACM International Conference on Multimedia*.

Zebang Cheng, Fuqiang Niu, Yuxiang Lin, Zhiqi Cheng, Xiaojiang Peng, and Bowen Zhang. 2024. MIPS at SemEval-2024 task 3: Multimodal emotion-cause pair extraction in conversations with multimodal language models. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 667–674, Mexico City, Mexico. Association for Computational Linguistics.

Hyung Won Chung, Le Hou, S. Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Wei Yu, Vincent Zhao, Yanping Huang, Andrew M. Dai, Hongkun Yu, Slav Petrov, Ed Huai hsin Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models. *ArXiv*, abs/2210.11416.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *North American Chapter of the Association for Computational Linguistics*.

Zixiang Ding, Huihui He, Mengran Zhang, and Rui Xia. 2019. From independent prediction to reordered prediction: Integrating relative position and global label information to emotion cause identification. In *AAAI Conference on Artificial Intelligence (AAAI)*, pages 6343–6350.

Zixiang Ding, Rui Xia, and Jianfei Yu. 2020a. ECPE-2D: Emotion-cause pair extraction based on joint two-dimensional representation, interaction and prediction. In *Association for Computational Linguistics (ACL)*, pages 3161–3170.

Zixiang Ding, Rui Xia, and Jianfei Yu. 2020b. End-to-end emotion-cause pair extraction based on sliding window multi-label learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3574–3583.

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2022. Glm: General language model pretraining with autoregressive blank infilling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 320–335.

P. Ekman. 1971. Universals and cultural differences in facial expressions of emotion. *Nebraska Symposium on Motivation. Nebraska Symposium on Motivation*, Vol. 19.

Paul Ed Ekman and Richard J Davidson. 1994. *The nature of emotion: Fundamental questions.* Oxford University Press.

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

Ana Ezquerro and David Vilares. 2024. LyS at SemEval-2024 task 3: An early prototype for end-to-end multimodal emotion linking as graph-based parsing. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 1252–1259, Mexico City, Mexico. Association for Computational Linguistics.

Hao Fei, Bobo Li, Qian Liu, Lidong Bing, Fei Li, and Tat seng Chua. 2023. Reasoning implicit sentiment with chain-of-thought prompting. In *Annual Meeting of the Association for Computational Linguistics*.

Qinghong Gao, Jiannan Hu, Ruifeng Xu, Gui Lin, Yulan He, Qin Lu, and Kam-Fai Wong. 2017. Overview of ntcir-13 eca task. In *Proceedings of the NTCIR-13 Conference*.

Alireza Ghahramani Kure, Mahshid Dehghani, Mohammad Mahdi Abootorabi, Nona Ghazizadeh, Seyed Arshan Dalili, and Ehsaneddin Asgari. 2024. AIMA at SemEval-2024 task 3: Simple yet powerful emotion cause pair analysis. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 1698–1703, Mexico City, Mexico. Association for Computational Linguistics.

Diman Ghazi, Diana Inkpen, and Stan Szpakowicz. 2015. Detecting emotion stimuli in emotion-bearing sentences. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 152–165. Springer.

Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. 2023. Imagebind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15180–15190.

Lin Gui, Dongyin Wu, Ruifeng Xu, Qin Lu, Yu Zhou, et al. 2016. Event-driven emotion cause extraction with corpus construction. In *EMNLP*, pages 1639–1649. World Scientific.

Hongyu Guo, Xueyao Zhang, Yiyang Chen, Lin Deng, and Binyang Li. 2024. UIR-ISC at SemEval-2024 task 3: Textual emotion-cause pair extraction in conversations. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 770–776, Mexico City, Mexico. Association for Computational Linguistics.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. *arXiv preprint arXiv:2111.09543*.

Chao-Chun Hsu, Sheng-Yeh Chen, Chuan-Chun Kuo, Ting-Hao Huang, and Lun-Wei Ku. 2018. Emotionlines: An emotion corpus of multi-party conversations. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdel rahman Mohamed. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460.

Srinivas Iyer, Xi Victoria Lin, Ramakanth Pasunuru, Todor Mihaylov, Daniel Simig, Ping Yu, Kurt Shuster, Tianlu Wang, Qing Liu, Punit Singh Koura, Xian Li, Brian O'Horo, Gabriel Pereyra, Jeff Wang, Christopher Dewan, Asli Celikyilmaz, Luke Zettlemoyer, and Veselin Stoyanov. 2022. Opt-iml: Scaling language model instruction meta learning through the lens of generalization. *ArXiv*, abs/2212.12017.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. 2019. Spanbert: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, 8:64–77.

Roman Kazakov, Kseniia Petukhova, and Ekaterina Kochmar. 2024. PetKaz at SemEval-2024 task 3: Advancing emotion classification with an LLM for emotion-cause pair extraction in conversations. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 1127–1134, Mexico City, Mexico. Association for Computational Linguistics.

Evgeny Kim and Roman Klinger. 2018. Who feels what and why? annotation of a literature corpus with semantic roles of emotions. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1345–1359.

Taewoon Kim and Piek Vossen. 2021. Emoberta: Speaker-aware emotion recognition in conversation with roberta. *ArXiv*, abs/2108.12009.

Sophia Yat Mei Lee, Ying Chen, and Chu-Ren Huang. 2010. A text-driven rule-based system for emotion cause detection. In *NAACL HLT Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, pages 45–53.

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

Sofiia Levchenko, Rafał Wolert, and Piotr Andruszkiewicz. 2024. PWEITINLP at SemEval-2024 task 3: Two step emotion cause analysis. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 1097–1105, Mexico City, Mexico. Association for Computational Linguistics.

Shu Li, Zicen Liao, and Huizhi Liang. 2024. NCL team at SemEval-2024 task 3: Fusing multimodal pre-training embeddings for emotion cause prediction in conversations. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 285–290, Mexico City, Mexico. Association for Computational Linguistics.

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

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. Dailydialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 986–995.

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

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024. Visual instruction tuning. *Advances in neural information processing systems*, 36.

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

Meng Luo, Han Zhang, Shengqiong Wu, Bobo Li, Hong Han, and Hao Fei. 2024. NUS-emo at SemEval-2024 task 3: Instruction-tuning LLM for multimodal emotion-cause analysis in conversations. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 1589–1596, Mexico City, Mexico. Association for Computational Linguistics.

Suyash Vardhan Mathur, Akshett Jindal, Hardik Mittal, and Manish Shrivastava. 2024. LastResort at SemEval-2024 task 3: Exploring multimodal emotion cause pair extraction as sequence labelling task. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 1204–1211, Mexico City, Mexico. Association for Computational Linguistics.

Trisha Mittal, Uttaran Bhattacharya, Rohan Chandra, Aniket Bera, and Dinesh Manocha. 2020. M3er: Multiplicative multimodal emotion recognition using facial, textual, and speech cues. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 1359–1367.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke E. Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Francis Christiano, Jan Leike, and Ryan J. Lowe. 2022. Training language models to follow instructions with human feedback. *ArXiv*, abs/2203.02155.

Victor Pacheco, Elizabeth Martinez, Juan Cuadrado, Juan Carlos Martinez Santos, and Edwin Puertas. 2024. VerbaNexAI lab at SemEval-2024 task 3: Deciphering emotional causality in conversations using multimodal analysis approach. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 1339–1343, Mexico City, Mexico. Association for Computational Linguistics.

Baolin Peng, Chunyuan Li, Pengcheng He, Michel Galley, and Jianfeng Gao. 2023. Instruction tuning with gpt-4. *ArXiv*, abs/2304.03277.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2019a. Meld: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 527–536.

Soujanya Poria, Navonil Majumder, Devamanyu Hazarika, Deepanway Ghosal, Rishabh Bhardwaj, Samson Yu Bai Jian, Pengfei Hong, Romila Ghosh, Abhinaba Roy, Niyati Chhaya, et al. 2021. Recognizing emotion cause in conversations. *Cognitive Computation*, pages 1–16.

Soujanya Poria, Navonil Majumder, Rada Mihalcea, and Eduard Hovy. 2019b. Emotion recognition in conversation: Research challenges, datasets, and recent advances. *IEEE Access*, 7:100943–100953.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for squad. *ArXiv*, abs/1806.03822.

Nicolay Rusnachenko and Huizhi Liang. 2024. nicolay-r at SemEval-2024 task 3: Using flan-t5 for reasoning emotion cause in conversations with chain-of-thought on emotion states. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 22–27, Mexico City, Mexico. Association for Computational Linguistics.

Irene Russo, Tommaso Caselli, Francesco Rubino, Ester Boldrini, and Patricio Martínez-Barco. 2011. Emocause: an easy-adaptable approach to emotion cause contexts. In *Workshop on Computational Approaches to Subjectivity and Sentiment Analysis (WASSA)*, pages 153–160.

Hidetsune Takahashi. 2024. Hidetsune at SemEval-2024 task 3: A simple textual approach to emotion classification and emotion cause analysis in conversations using machine learning and next sentence prediction. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 361–364, Mexico City, Mexico. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin R. Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Daniel M. Bikel, Lukas Blecher, Cristian Cantón Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony S. Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel M. Kloumann, A. V. Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan

Schelten, Ruan Silva, Eric Michael Smith, R. Subramanian, Xia Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zhengxu Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *ArXiv*, abs/2307.09288.

Fanfan Wang, Zixiang Ding, Rui Xia, Zhaoyu Li, and Jianfei Yu. 2023a. Multimodal emotion-cause pair extraction in conversations. *IEEE Transactions on Affective Computing*, 14(3):1832–1844.

Fanfan Wang, Jianfei Yu, and Rui Xia. 2023b. Generative emotion cause triplet extraction in conversations with commonsense knowledge. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 3952–3963.

Zining Wang, Yanchao Zhao, Guanghui Han, and Yang Song. 2024. QFNU_CS at SemEval-2024 task 3: A hybrid pre-trained model based approach for multimodal emotion-cause pair extraction task. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 349–353, Mexico City, Mexico. Association for Computational Linguistics.

Rui Xia and Zixiang Ding. 2019. Emotion-cause pair extraction: A new task to emotion analysis in texts. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1003–1012.

Rui Xia, Mengran Zhang, and Zixiang Ding. 2019. RTHN: A RNN-transformer hierarchical network for emotion cause extraction. In *International Joint Conference on Artificial Intelligence (IJCAI)*, pages 5285–5291.

Zhengyuan Yang, Linjie Li, Kevin Lin, Jianfeng Wang, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. 2023. The dawn of lmms: Preliminary explorations with gpt-4v (ision). *arXiv preprint arXiv:2309.17421*, 9(1):1.

Duzhen Zhang, Zhen Yang, Fandong Meng, Xiuyi Chen, and Jie Zhou. 2022. Tsam: A two-stream attention model for causal emotion entailment. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6762–6772.

Hang Zhang, Xin Li, and Lidong Bing. 2023. Video-llama: An instruction-tuned audio-visual language model for video understanding. *ArXiv*, abs/2306.02858.

Shen Zhang, Haojie Zhang, Jing Zhang, Xudong Zhang, Yimeng Zhuang, and Jinting Wu. 2024. Samsung research China-Beijing at SemEval-2024 task 3: A multi-stage framework for emotion-cause pair extraction in conversations. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 536–546, Mexico City, Mexico. Association for Computational Linguistics.

Jinming Zhao, Tenggan Zhang, Jingwen Hu, Yuchen Liu, Qin Jin, Xinchao Wang, and Haizhou Li. 2022. M3ed: Multi-modal multi-scene multi-label emotional dialogue database. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5699–5710.

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.