# NUS-Emo at SemEval-2024 Task 3: Instruction-Tuning LLM for Multimodal Emotion-Cause Analysis in Conversations

**Meng Luo**[1*] **Han Zhang**[2*] **Shengqiong Wu**[1] **Bobo Li**[3] **Hong Han**[2] **Hao Fei**[1†]
[1]National University of Singapore [2]Xidian University [3]Wuhan University
mluo@u.nus.edu zhanghanxd@stu.xidian.edu.cn swu@u.nus.edu
boboli@whu.edu.cn hanh@mail.xidian.edu.cn haofei37@nus.edu.sg

## Abstract

This paper describes the architecture of our system developed for Task 3 of SemEval-2024: Multimodal Emotion-Cause Analysis in Conversations. Our project targets the challenges of subtask 2, dedicated to Multimodal Emotion-Cause Pair Extraction with Emotion Category (MECPE-Cat), and constructs a dual-component system tailored to the unique challenges of this task. We divide the task into two subtasks: emotion recognition in conversation (ERC) and emotion-cause pair extraction (ECPE). To address these subtasks, we capitalize on the abilities of Large Language Models (LLMs), which have consistently demonstrated state-of-the-art performance across various natural language processing tasks and domains. Most importantly, we design an approach of emotion-cause-aware instruction-tuning for LLMs, to enhance the perception of the emotions with their corresponding causal rationales. Our method enables us to adeptly navigate the complexities of MECPE-Cat, achieving a weighted average 34.71% F1 score of the task, and securing the $\mathbf{2}^{nd}$ rank on the leaderboard.[1] The code and metadata to reproduce our experiments are all made publicly available.[2]

## 1 Introduction

Emotion cause analysis is a critical component of human communication and decision-making, offering substantial applications across diverse fields. It enables a deeper and more detailed understanding of sentiments. The introduction of emotion-cause analysis in textual conversations by Poria et al. (2021); Xia and Ding (2019) has paved the way for advancements in understanding emotional dynamics within dialogues. However, textual analysis alone does not fully capture the complexity of human emotional expression, as emotions and their causes are often conveyed through a blend of modalities (Hazarika et al., 2018; Wu et al., 2023a; Fei et al., 2023b). Subtask 2 of SemEval-2024 Task 3, referred to as MECPE-Cat, seeks to expand this analysis into the multimodal domain, focusing on English-language conversations. The task draws inspiration from the seminal work of Wang et al. (2023), which sets out to jointly extract emotions and their corresponding causes from conversations across multiple modalities, including text, audio, and video, and it also encompasses the identification of the corresponding emotion category for each emotion-cause pair.

In our system, we leverage LLMs such as GPT-3 (Brown et al., 2020), Flan-T5 (Chung et al., 2022), and GLM (Du et al., 2021) known for their exceptional performance in various natural language processing tasks. We employ parameter-efficient fine-tuning, specifically LoRA (Hu et al., 2021), to efficiently fine-tune LLMs, enhancing their performance with minimal computational overhead. Additionally, we harness emotion-cause-aware prompt-based learning and instruction-tuning to enhance model performance such that the LLMs can more accurately perceive the emotions with their corresponding causal rationales. Prompt-based learning guides LLMs to generate contextually relevant outputs, while instruction-fine-tuning models for our specific tasks by improving their response to explicit instructions.

In this paper, we investigate the optimal LLM for the MECPE-Cat task, selecting ChatGLM based on its superior zero-shot performance. We further refine ChatGLM through instruction-tuning, using carefully crafted prompts to enhance its task-specific accuracy. Our fine-tuned model achieves the second-highest score on the official test set for subtask 2, with a weighted average of 34.71% F1,

[*]Equal contributions.
[†]Corresponding author.
[1]https://nustm.github.io/SemEval-2024_ECAC/
[2]https://github.com/zhanghanXD/NUS-Emo-at-SemEval-2024-Task3

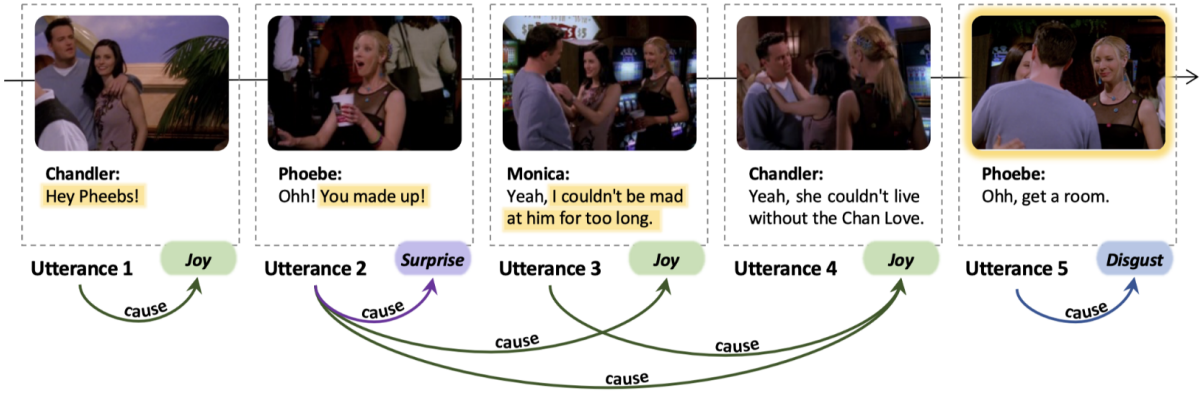


Figure 1: An example of an official task and annotated dataset. Each arc points from the cause utterance to the emotional triggers. The cause spans have been highlighted in yellow. Background: Chandler and his girlfriend Monica walked into the casino (they had a quarrel earlier but made up soon), and then started a conversation with Phoebe.

underscoring the effectiveness of our approach. We also discuss the current limitations of our model and methodology, alongside directions for future research and improvement. We will release our codes and resources mentioned in this paper to facilitate relevant research.

## 2 Background

### 2.1 Task and Dataset Description

The SemEval-2024 Task 3 (Wang et al., 2024) is based on the multimodal conversational emotion-cause dataset, Emotion-Cause-in-Friends (ECF; Wang et al., 2023), by choosing a multimodal dataset MELD (Poria et al., 2018) as the data source and further annotating the corresponding causes for the given emotion annotations. The ECF dataset contains 9,794 emotion-cause pairs, covering three modalities. The subtask 2 is to extract all emotion-cause pairs in a given conversation under three modalities, where each pair contains an emotion utterance along with its emotion category and a cause utterance, e.g., (U3_Joy, U2), which means that the speaker's joy emotion in utterance 3 is triggered by the cause from utterance 2. Figure 1 displays a real example of this task and annotated dataset. In this conversation, it is expected to extract a set of six utterance-level emotion-cause pairs in total, e.g., Chandler's Joy emotion in Utterance 4 (U4 for short) is triggered by the objective cause that he and Monica had made up and Monica's subjective opinion in U3, forming the pairs (U4_joy, U2) and (U4_joy, U3); The cause for Phoebe's Disgust in U5 is the objective event that Monica and Chandler were kissing in front of her (mainly reflected in the visual modality of U5), forming the pair (U5_disgust, U5).

### 2.2 Related Work

The exploration of ECPE within textual and conversational contexts has been approached through various methodologies, each tailored to specific task settings (Chen et al., 2022). Cheng et al. (2023) reframe the ECPE task as a process akin to engaging in a two-stage machine reading comprehension (MRC) challenge. Zheng et al. (2023) expand the ECPE task to Emotion-Cause Quadruple Extraction in Dialogs (ECQED), focusing on detecting pairs of emotion-cause utterances and their types. They present a model utilizing a heterogeneous graph and a parallel grid tagging scheme for this purpose. In addressing the specific challenge of the MECPE-Cat task, Wang et al. (2023) set a benchmark for this task by introducing two preliminary baseline systems. They utilize a heuristic approach to leverage inherent patterns in the localization of causes and emotions, alongside a deep learning strategy, MECPE-2steps, which adapts a prominent ECPE methodology for news articles to include multimodal data.

Drawing from the varied methodologies of previous work, it becomes clear that effectively solving the MECPE-Cat task demands a deep understanding of dialogue content, precise identification of conversational emotions, extraction of emotion-cause pairs, and the integration of multimodal information. Motivated by the strong performance of LLMs on various metrics, we opt to utilize these models to address this intricate challenge. Through exhaustive model evaluations and extensive prompt

testing, we have showcased the practicality, superiority, and adaptability of our chosen approach.

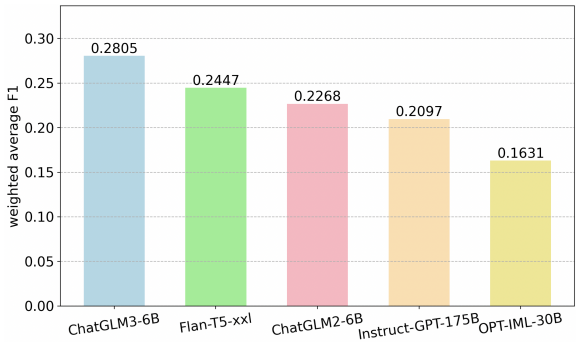


Figure 2: Zero-shot test set performance of various instruction-tuned LLMs.

## 3 Methodology

In this section, we first conduct preliminary experiments to determine which LLM to select as a backbone reasoner. We then elaborate on how we design the system and emotion-cause-aware instructions for tuning our chosen LLM.

### 3.1 Pilot Study for LLM Selection

Currently, there exists a variety of LLMs, such as OPT-IML (Iyer et al., 2022), GPT-3, Flan-T5, and GLM. However, it is essential to select a model that not only performs optimally but is also the most suitable for our specific task. To this end, we carry out a pilot study to determine the most appropriate model selection. For our zero-shot testing experiment, we rigorously evaluate several models, including OPT-IML[3], Instruct-GPT[4] (Ouyang et al., 2022), Flan-T5[5], alongside the ChatGLM models, to identify the most effective tool for this specific task. We customize instructions for each model's specific tuning style, recognizing that a single set of instructions does not suit all models effectively. We also embed expected output labels within these instructions to secure precise responses from each model. Figure 2 depicts the zero-shot performance of these models. The ChatGLM[6] LLM is ultimately selected based on its superior performance in these tests. This selection is informed not merely by the innovative features or the advanced training methodologies of ChatGLM but by empirical evidence of its exceptional zero-shot performance among the models considered.

[3]OPT-IML-30B, max version with 30B, https://huggingface.co/facebook/opt-iml-30b

[4]Instruct-GPT-175B, an advanced version of the GPT-3.5.

[5]Flan-T5-xxl, with 11B, https://huggingface.co/google/flan-t5-xxl

[6]ChatGLM, 3rd version with 6B, https://github.com/THUDM/ChatGLM3.

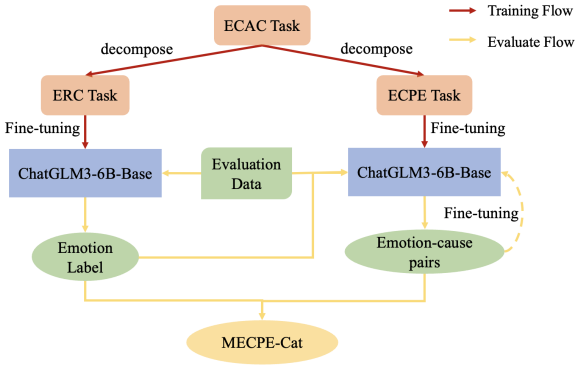


Figure 3: Proposed method workflow for the MECPE-Cat task.

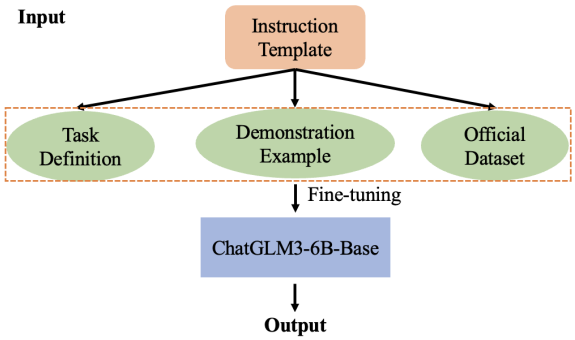


Figure 4: The construction of the instruction template and the flow of model input and output.

### 3.2 Multimodal Feature Encoding

Given that the inputs for our task incorporate multimodal signals, including visual information to assist in more accurate emotion recognition, it is imperative to fully leverage the non-textual modal information. However, our LLM backbone does not natively support the direct inclusion of non-textual modal signals. To address this, we consider employing ImageBind (Girdhar et al., 2023) for encoding the multimodal portion of input information, owing to its robust multimodal alignment capabilities and visual perception proficiency. Subsequently, we concatenate the multimodal representations with other textual embeddings before feeding them into the LLM.

### 3.3 Constructing Emotion-Cause-aware Instructions for LLM Tuning

Figure 3 first illustrates the workflow of our proposed framework. Initially, we fine-tune the model on the ERC task. Following this, we incorporate

the predicted emotion labels into each utterance, setting the stage for the ECPE task execution. Subsequently, we employ the model, now fine-tuned with data labeled with emotion tags, to perform inference on the MECPE-Cat task, yielding an initial set of emotion-cause pairs. These preliminary results are then reintegrated into the original training dataset for a second round of fine-tuning, culminating in the refinement of our model to produce the final set of emotion-cause pairs.

**Task Definition:**
*"You're an expert in sentiment analysis and emotion cause identification. Below is a conversation containing multiple utterances from different speakers, along with the corresponding emotion label for each utterance. Your task is to identify the indices of the candidate utterances that elicited the emotion in the target utterance."*

**Input conversation:**
*1_joy. Chandler: Hey Pheebs!*
*2_surprise. Phoebe: Ohh! You made up!*
*3_joy. Monica: Yeah, I couldn't be mad at him for too long.*
*4_joy. Chandler: Yeah, she couldn't live without the Chan Love.*
*5_disgust. Phoebe: Ohh, get a room.*

**Candidate utterances:**
*1_joy. Chandler: Hey Pheebs!*
*2_surprise. Phoebe: Ohh! You made up!*
*3_joy. Monica: Yeah, I couldn't be mad at him for too long.*

**Target utterance:**
*4_joy. Chandler: Yeah, she couldn't live without the Chan Love.*

**Question:**
*The emotion-cause indices of the target utterance are:*
[LLM output]

To enhance the perception of identifying emotion-cause pairs and mitigate the task's inherent complexity and potential confusion, we design the template for producing emotion-cause-aware instructions to guide the model. Figure 4 illustrates the construction of the instruction template, which encompasses the task definition, a demonstration example, and the dataset for which the model is expected to predict outcomes. This structured approach not only simplifies the task's complexity for the model but also aligns the model's processing capabilities with the requirements of accurately identifying emotion-cause pairs in conversations. In the above box we showcase a real example.

## 4 Experiments

This section will quantify the effectiveness of our systems via experiments and also show more analyses to gain more observations.

### 4.1 Implementation

The hyperparameter of our system used to achieve the highest weighted average F1 score on the subtask 2 is listed in 1. The ChatGLM model was fine-tuned using a learning rate of 1e-4 with LoRA-specific configurations including a rank of 8, alpha value of 32, and a dropout rate of 0.1. The training was conducted with a maximum instruction length of 2048 tokens and an output length limited to 128 tokens, using a batch size of 1. We used a single gradient accumulation step across 2 training epochs. These parameters were meticulously selected to optimize our model's performance.

| Hyperparameter | Value |
|---|---|
| Learning rate | 1e-4 |
| LoRA rank | 8 |
| LoRA alpha | 32 |
| LoRA dropout | 0.1 |
| Max instruction length | 2048 |
| Max output length | 128 |
| Batch size | 1 |
| Gradient accumulation steps | 1 |
| Epochs | 2 |

Table 1: Hyperparameter used for the best performing model.

### 4.2 Evaluating Template Designing

In constructing the instruction dataset for tuning LLMs, we systematically transform each dialogue in the dataset into training samples by embedding them into a fixed template as described above. The data source for this transformation is the officially provided ECF dataset, which comprises 13,619 utterances. Consequently, we constructed a total of 13,619 templates based on this dataset, each

| Condition | F1 Score |
|---|---|
| Only Task Definition | 0.2981 |
| Task + Example | 0.3124 |
| Task + Example + Candidate | **0.3207** |

Table 2: Performance using different templates for constructing instruction tuning.

tailored to facilitate the model's learning and application of emotion-cause-aware instructions.

We here perform an ablation study on the contributions of each part of the instructions we designed for the task. We derive three variants:

- **Only Task Definition:** Compared to the zero-shot paradigm, this condition offers a more detailed and precise description of the task.
- **Task + Example:** We provide a demonstrative example to clearly show the expected outcome in a real-world dialogue, offering the model a practical reference for task execution
- **Task + Example + Candidate utterances:** This design simplifies the task by introducing 'candidate utterances,' enabling the model to analyze emotion-cause pairs sentence by sentence, rather than across entire dialogues, and pinpoint the specific causes of emotions from the preceding content.

Table 2 demonstrates the comparative performance of these diverse templates. We see that different components of the instruction templates show clear influences, such as task definition, example demonstration, and candidate utterances. Thus, we apply all these components into our instruction templates.

### 4.3 Instruction-tuning LLM

For our experiments, we adopt a meticulous fine-tuning process for the ChatGLM. We set a learning rate of 1e-4, aiming for a balance between rapid convergence and maintaining the model's ability to adapt without overfitting. We leverage the LoRA technique with a rank of 8 and alpha of 32 to introduce task-adaptive parameters without bloating the model size, alongside a dropout rate of 0.1 to prevent overfitting. The model processed inputs with a max sequence length of 2048 tokens, accommodating the depth of context required for our task, while the outputs are capped at 128 tokens to focus on generating concise and relevant responses. Both batch size and gradient accumulation steps are set to 1, tailored to our computational resources while ensuring effective backpropagation. This configuration, selected after careful evaluation of various setups, is instrumental in fine-tuning the ChatGLM model to achieve the best performance on our task.

Our experiments capitalize on the robust computational capabilities provided by NVIDIA A800-SXM GPUs, each boasting 80 GB of VRAM, to ensure sufficient resources are available to train large language models. This fine-tuning process is facilitated using a customized script derived from the Hugging Face Transformers framework, chosen for its extensive support of transformer models and seamless integration with our setup, thereby enabling us to leverage advanced hardware capabilities while utilizing a leading-edge software environment for our model's optimization.

### 4.4 Task Decomposition

We decompose the MECPE-Cat task into ERC and ECPE phases to strategically alleviate its complexity. This division offered a two-fold advantage: firstly, it distills the task into clearer, more focused components, facilitating a more straightforward understanding and execution of the model. Secondly, by leveraging emotion labels obtained from the ERC phase during the ECPE phase, we enhance the model's capability to pinpoint emotion-cause pairs with greater accuracy. Tabel 3 showcases incremental improvements in weighted average F1 scores across three distinct setups. This progression underscores the dual benefits of our approach: simplifying the task's complexity for the model and enriching the ECPE phase with contextual emotion labels, thereby optimizing the extraction of emotion-cause pairs.

| Methods | F1 Score |
|---|---|
| Single Stage | 0.3207 |
| Two Independent Stages | 0.3288 |
| ECPE with Emotion Labels | **0.3396** |

Table 3: Comparison of weighted average F1 Scores under different methods.

### 4.5 Data Augmentation

We find that augmenting the training dataset with trial data significantly enhanced model accuracy, achieving a high weighted average F1 score of 0.3416, as shown in Table 4. Furthermore, we employ a trick by incorporating the model's inference results on the ECPE task back into the training dataset for an additional round of fine-tuning. This iterative fine-tuning strategy yielded a further improvement in our test data performance. These

enhancements demonstrate the efficacy of not only expanding the training dataset but also utilizing the model's own outputs to refine its accuracy.

| Data | Epoch 1 | Epoch 2 | Epoch 3 |
|---|---|---|---|
| Train | 0.3390 | 0.3396 | 0.3393 |
| Train + Trial | 0.3404 | 0.3410 | 0.3406 |
| Iterative Train | 0.3408 | **0.3416** | 0.3411 |

Table 4: Comparison of weighted average F1 Scores across different training data and epochs.

### 4.6 Multimodal Integration

To assess the impact of multimodal information on our model's performance, we adopt a methodological approach that harnessed GPT-4V Achiam et al. (2023) for extracting insights from modalities beyond text. Specifically, we enrich the instruction template with "video description of target utterance" derived from GPT-4V, presenting it as supplementary information to guide the model. This strategic integration of multimodal data leads to an improvement in the model's F1 score, as shown in Table 5, which validates the utility of multimodal information in providing richer contextual understanding.

| Information | F1 Score |
|---|---|
| Text | 0.3416 |
| Text + Video | **0.3471** |

Table 5: Comparison of weighted average F1 Scores between pure text and multimodal information.

## 5 Conclusion

In this work, we explore the LLMs for solving the Multimodal Emotion-Cause Pair Extraction with Emotion Category (MECPE-Cat) task. Through a pilot study, we first select an LLM, ChatGLM, that assists in achieving optimal task performance. The backbone ChatGLM receives textual dialogue, and also perceives the multimodal information via the ImageBind vision encoder. Lastly, we devise an emotion-cause-aware instruction-tuning mechanism for updating LLMs, which enhances the perception of the emotions with their corresponding causal rationales. Our system achieves a weighted average F1 score of 34.71%, securing second place on the MECPE-Cat leaderboard.

## Acknowledgements

This work is sponsored by CCF-Baidu Open Fund.

## References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Yuyang Chai, Chong Teng, Hao Fei, Shengqiong Wu, Jingye Li, Ming Cheng, Donghong Ji, and Fei Li. 2022. Prompt-based generative multi-label emotion prediction with label contrastive learning. In *CCF International Conference on Natural Language Processing and Chinese Computing*, pages 551–563. Springer.

Shunjie Chen, Xiaochuan Shi, Jingye Li, Shengqiong Wu, Hao Fei, Fei Li, and Donghong Ji. 2022. Joint alignment of multi-task feature and label spaces for emotion cause pair extraction. In *Proceedings of the 29th International Conference on Computational Linguistics, COLING 2022, Gyeongju, Republic of Korea, October 12-17, 2022*, pages 6955–6965.

Zifeng Cheng, Zhiwei Jiang, Yafeng Yin, Cong Wang, Shiping Ge, and Qing Gu. 2023. A consistent dual-mrc framework for emotion-cause pair extraction. *ACM Transactions on Information Systems*, 41(4):1–27.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2021. Glm: General language model pretraining with autoregressive blank infilling. *arXiv preprint arXiv:2103.10360*.

Hao Fei, Bobo Li, Qian Liu, Lidong Bing, Fei Li, and Tat-Seng Chua. 2023a. Reasoning implicit sentiment with chain-of-thought prompting. *arXiv preprint arXiv:2305.11255*.

Hao Fei, Qian Liu, Meishan Zhang, Min Zhang, and Tat-Seng Chua. 2023b. Scene graph as pivoting: Inference-time image-free unsupervised multimodal machine translation with visual scene hallucination. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 5980–5994.

Hao Fei, Yafeng Ren, Yue Zhang, and Donghong Ji. 2023c. Nonautoregressive encoder-decoder neural framework for end-to-end aspect-based sentiment

triplet extraction. *IEEE Trans. Neural Networks Learn. Syst.*, 34(9):5544–5556.

Hao Fei, Yue Zhang, Yafeng Ren, and Donghong Ji. 2020. Latent emotion memory for multi-label emotion classification. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 7692–7699.

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.

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

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

Srinivasan Iyer, Xi Victoria Lin, Ramakanth Pasunuru, Todor Mihaylov, Dániel Simig, Ping Yu, Kurt Shuster, Tianlu Wang, Qing Liu, Punit Singh Koura, et al. 2022. Opt-iml: Scaling language model instruction meta learning through the lens of generalization.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. *arXiv preprint arXiv:2104.08691*.

Bobo Li, Hao Fei, Fei Li, Yuhan Wu, Jinsong Zhang, Shengqiong Wu, Jingye Li, Yijiang Liu, Lizi Liao, Tat-Seng Chua, et al. 2022. Diaasq: A benchmark of conversational aspect-based sentiment quadruple analysis. *arXiv preprint arXiv:2211.05705*.

Bobo Li, Hao Fei, Lizi Liao, Yu Zhao, Chong Teng, Tat-Seng Chua, Donghong Ji, and Fei Li. 2023. Revisiting disentanglement and fusion on modality and context in conversational multimodal emotion recognition. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 5923–5934.

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

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

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

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*, 13:1317–1332.

Leigang Qu, Shengqiong Wu, Hao Fei, Liqiang Nie, and Tat-Seng Chua. 2023. Layoutllm-t2i: Eliciting layout guidance from llm for text-to-image generation. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 643–654.

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

Fanfan Wang, Heqing Ma, Rui Xia, Jianfei Yu, and Erik Cambria. 2024. Semeval-2024 task 3: Multimodal emotion cause analysis in conversations. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 2022–2033, Mexico City, Mexico. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pages 38–45.

Shengqiong Wu, Hao Fei, Wei Ji, and Tat-Seng Chua. 2023a. Cross2stra: Unpaired cross-lingual image captioning with cross-lingual cross-modal structure-pivoted alignment. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 2593–2608. Association for Computational Linguistics.

Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. 2023b. Next-gpt: Any-to-any multimodal llm. *arXiv preprint arXiv:2309.05519*.

Shengqiong Wu, Hao Fei, Hanwang Zhang, and Tat-Seng Chua. 2024. Imagine that! abstract-to-intricate text-to-image synthesis with scene graph hallucination diffusion. *Advances in Neural Information Processing Systems*, 36.

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.

Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, et al. 2022. Glm-130b: An open bilingual pre-trained model. *arXiv preprint arXiv:2210.02414*.

Ao Zhang, Hao Fei, Yuan Yao, Wei Ji, Li Li, Zhiyuan Liu, and Tat-Seng Chua. 2024a. Vpgtrans: Transfer visual prompt generator across llms. *Advances in Neural Information Processing Systems*, 36.

Meishan Zhang, Bin Wang, Hao Fei, and Min Zhang. 2024b. In-context learning for few-shot nested named entity recognition. *arXiv preprint arXiv:2402.01182*.

Shengyu Zhang, Linfeng Dong, Xiaoya Li, Sen Zhang, Xiaofei Sun, Shuhe Wang, Jiwei Li, Runyi Hu, Tianwei Zhang, Fei Wu, et al. 2023. Instruction tuning for large language models: A survey. *arXiv preprint arXiv:2308.10792*.

Yu Zhao, Hao Fei, Yixin Cao, Bobo Li, Meishan Zhang, Jianguo Wei, Min Zhang, and Tat-Seng Chua. 2023. Constructing holistic spatio-temporal scene graph for video semantic role labeling. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 5281–5291.

Li Zheng, Donghong Ji, Fei Li, Hao Fei, Shengqiong Wu, Jingye Li, Bobo Li, and Chong Teng. 2023. Ecqed: Emotion-cause quadruple extraction in dialogs. *arXiv preprint arXiv:2306.03969*.