DeepPavlov at SemEval-2024 Task 3: Multimodal Large Language Models in Emotion Reasoning

Julia Belikova and **Dmitrii Kosenko**
Moscow Institute of Physics and Technology
belikova.iua@phystech.edu, kosenko.dp@mipt.ru

Abstract

This paper presents the solution of the DeepPavlov team for the Multimodal Sentiment Cause Analysis competition in SemEval-2024 Task 3, Subtask 2 (Wang et al., 2024). In the evaluation leaderboard, our approach ranks 7th with an F1-score of 0.2132. Large Language Models (LLMs) are transformative in their ability to comprehend and generate human-like text. With recent advancements, Multimodal Large Language Models (MLLMs) have expanded LLM capabilities, integrating different modalities such as audio, vision, and language. Our work delves into the state-of-the-art MLLM Video-LLaMA, its associated modalities, and its application to the emotion reasoning downstream task, Multimodal Emotion Cause Analysis in Conversations (MECAC). We investigate the model's performance in several modes: zero-shot, few-shot, individual embeddings, and fine-tuned, providing insights into their limits and potential enhancements for emotion understanding.

1 Introduction

In the dynamic domain of artificial intelligence, the emergence of MLLMs has gained significant interest due to integrating input from different modalities, such as audio, vision and language, opens up in-depth perceptual and interpretive capabilities in dialogues instead of chat-based dialogue systems (Konovalov et al., 2016).

These models exhibit impressive potential in resolving a lot of challenges and have been deployed across various sectors, including banking support systems, social services, and as adjuncts in psychological assistance. In these applications, the deciphering of user intent and emotions is crucial for generating pertinent responses. Consequently, one of the most important domains within MLLM research is emotion reasoning.

This paper explores a specific facet of emotion reasoning: Multimodal Emotion Cause Analysis in Conversations (MECAC) (Wang et al., 2024). This task involves emotion recognition and matching emotional states with their causes in the context of a conversation, leveraging inputs from different modalities such as text, audio, video and more.

Despite the remarkable advancements in MLLMs, their capabilities and limitations in the high-potential area of emotion reasoning remain active topics for research. One of the seminal contributions to the investigation of MLLMs capabilities within the area of emotional reasoning is delineated in (Lian et al., 2023). The authors introduce a novel task, Explainable Multimodal Emotion Reasoning (EMER), and proceed to evaluate the efficacy of modern multimodal models in addressing EMER. Their research focuses on the integration and interpretability of emotional cues across diverse modalities, thereby advancing the understanding of emotion reasoning.

To address the described issue, we propose to continue research of the MLLMs capabilities in emotion reasoning by evaluating one of the most promising models, Video-LLaMA (Zhang et al., 2023), also explored in Lian et al. (2023), for MECAC on the Emotion-Cause-in-Friends dataset.

The work evaluates the performance of the model in three modes:

1. *Zero-shot and Few-shot modes*. These modes are utilized to evaluate the model's initial capabilities in emotion reasoning.

2. *Individual Embeddings mode*. In this mode, embeddings from individual modalities are employed alongside trained basic heads to address MECAC.

3. *Fine-tuned mode*. This mode is used to evaluate specialized emotion reasoning capabilities.

Experimental results demonstrate the considerable potential of MLLMs in the domain of emotion reasoning.

Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024), pages 1747–1757

2 Related Work

2.1 Multimodal Large Language Models

The ascent of Large Language Models such as LLaMA2 (Touvron et al., 2023), Qwen (Bai et al., 2023), Mistral (Jiang et al., 2023) has marked a significant milestone in the field of artificial intelligence. These models have demonstrated exceptional capabilities in language reasoning and decision-making, closely mirroring human-level performance.

The integration of adapters to align pre-trained encoders from different modalities with textual LLMs has given rise to a new class of MLLMs such as: Flamingo (Alayrac et al., 2022), BLIP-2 (Li et al., 2023a), MiniGPT-4 (Zhu et al., 2023), mPLUG-Owl (Ye et al., 2023), VideoChat (Li et al., 2024a), InstructBLIP (Dai et al., 2023), Video-ChatGPT (Maaz et al., 2023), Video-LLaVA (Lin et al., 2023), VideoChat2 (Li et al., 2024b), Video-LLaMA (Zhang et al., 2023) (Table 1).

Model	Modality
Flamingo	I, V, T
BLIP-2	I, T
MiniGPT-4	I, T
mPLUG-Owl	I, V, T
InstructBLIP	I, T
Video-ChatGPT	I, V, T
VideoChat2	I, V, T
Video-LLaMA	I, V, A, T
Video-LLaVA	I, V, T

Table 1: Multimodal Large Language Models. T, A, I, and V stand for text, audio, image and silent video, respectively

These models have gained impressive results in well-known general domains (Wu et al., 2023): temporal perception and reasoning, casual inference, and spatial perception and analysis.

2.2 Emotion Reasoning in Conversation

Our work delves into MECAC, a derivation of ECPE (Xia and Ding, 2019), the downstream task of emotion reasoning. Given a conversation sequence consisting of N utterances, $U = \{U_1, U_2, \ldots, U_N\}$, where each utterance U_i is accompanied by a corresponding speaker identity, textual content, and an associated audio-visual clip.

The task is to output a set of emotion-cause pairs $E = \{(e_i, c_i)\}_{i=1}^{M}$, where each pair contains:

- e_i: an emotion utterance U_j that expresses an emotion.
- c_i: a cause utterance U_k that is identified as the cause of the emotion expressed in U_j.

Additionally, each emotion utterance e_i is tagged with an emotion category EC from a predefined set of emotion categories $EC = \{EC_1, EC_2, \ldots, EC_K\}$.

The exploration of emotion reasoning within the context of conversations has traditionally been addressed using various classical approaches. Recurrence-based or graph-based methods have been particularly popular due to their ability to capture sequential and relational data effectively. Notable methods in this domain include: MC-ECPE-2steps (Wang et al., 2023), which focuses on two-step recurrence-based emotion-cause pair extraction; Joint-GCN (Li et al., 2023b), which leverages recurrent and graph convolutional networks for joint emotion-cause detection; ECQED (Zheng et al., 2023), which extends the emotion-cause pair extraction to a quadruple extraction task and structural and semantic heterogeneous graph for conversation representation; CORECT (Nguyen et al., 2023), which enhances conversation understanding through relational temporal graph neural networks; and COGMEN (Joshi et al., 2022), which utilizes contextualized graph neural networks for multimodal emotion recognition.

In this paper, we investigate the capabilities of MLLMs for solving MECAC by evaluating the state-of-the-art model, Video-LLaMA. It is pertinent to acknowledge the application of MLLMs in a variety of sentiment and emotion recognition (Aslam et al., 2023), in the domain of EMER. Previous works indicate that MLLMs demonstrate notable efficacy in these complex tasks, which highlights their potential for advancing the frontier of emotion reasoning research.

2.3 Video-LLaMA architecture

Next, we summarize the key points of Video-LLaMA's architecture.

Video-LLaMA is a multimodal framework designed to extend the capabilities of frozen LLMs by enabling them to process and respond to audio-visual content.

Visual Encoder. The visual encoding component employs a pre-trained image encoder to compute representations from individual frames of a video. It introduces a frame embedding layer to

provide temporal information and incorporates a video Q-former to generate visual query tokens that encapsulate the temporal dynamics of visual scenes. A linear layer is introduced to transform video embedding vectors into query vectors that are compatible with the embedding space of LLMs.

Audio Encoder. For audio processing, Video-LLaMA leverages ImageBind (Girdhar et al., 2023). It also uses a similar architecture to the visual encoder to obtain audio embeddings for the LLM module.

Cross-Modal Training. The training process involves multi-branch, cross-modal pre-training to achieve both vision-language and audio-language alignment. The vision-language pre-training includes a video-clips-to-text generation task and static image-caption learning. The audio-language pre-training leverages the audio encoder and vision-text data to align with the LLM's embedding space.

Standard Inference. During inference, Video-LLaMA is capable of zero-shot video and audio understanding. It processes video frames and audio signals, converts them into query representations that are concatenated with textual input embeddings of LLMs, and generates responses grounded in the video's visual and auditory content.

3 Methods

3.1 Zero-shot and Few-shot

Today's LLMs are developed using extensive datasets and are further fine-tuned to comprehend and follow instructions, granting them the capacity to perform certain tasks in a zero-shot fashion (Tirskikh and Konovalov, 2023). Investigating how these capabilities are exhibited in multimodal models represents an active area of research.

To evaluate the capabilities of Video-LLaMA in the zero-shot emotion reasoning subtasks, we use structured templates such as the one detailed in Appendix A, Listing 4

While LLMs demonstrate remarkable zero-shot capabilities, they still fall short on more complex tasks within the zero-shot setting. Consequently, it is essential to evaluate their few-shot capabilities as well. To achieve this, we employ prompt templates, such as the one described in Appendix A, Listing 5.

3.2 Individual Embeddings

In this work, we also investigate the capabilities of embeddings obtained from the output of Video-LLaMA. Specifically, we extract multimodal embeddings corresponding to each fragment of the conversation. These embeddings are derived from the last semantic token of the last hidden state during the generation of responses to prompts formatted as shown in Listing 1.

```
# First option
<Item Value> Describe the behavior of
the speaker in this <Item Name> in one
word:
# Second option
<Item Value> Describe what is happening
in this <Item Name> in one word:
# Third option
<Item Value> Describe the emotional
state of the speaker in this <Item Name>
in one word:
```

Listing 1: Prompt templates for multimodal embeddings generation

To utilize these embeddings for our task, we integrate classical heads based on Multi-Layer Perceptron (MLP), Bidirectional Long Short-Term Memory (BiLSTM), and Self-Attention mechanisms. In the context of MLP, we adopt a straightforward approach for multi-class classification of emotions and binary classification of causes. Importantly, for the binary classification of causes, we consider all possible pairs of utterances. The probability that a pair belongs to a specific class is computed based on the output from the linear layer, which receives a concatenated representation of the utterance pairs as its input.

The BiLSTM head is implemented similarly to the MLP. For the self-attention mechanism, we employ multiple layers of a classical architecture. For both approaches — the MLP and BiLSTM-based heads — we utilize Cross-Entropy as the loss function.

It is also worth noting that there is a class imbalance in the case of binary causes classification. According to the authors of the dataset about 55.73% of the utterances are annotated with one of the six basic emotions, and 91.34% of the emotions are annotated with the corresponding causes in the ECF dataset. As a result, the matrices of some conversations divided into utterances become quite sparse. To mitigate the impact of imbalance, we propose several balancing methods: simple weighting of the loss function and adaptive weighting. In the first case, a constant scaling factor is chosen to increase the influence of the minority class, while in the

second case, balancing is done for each individual batch based on the current class frequency.

3.3 Fine-tuning

The fine-tuning stage employs Low-Rank Adaptation (LoRA) (Hu et al., 2021) to modify the pre-trained parameters of the LLaMA module within Video-LLaMA, while the visual and audio encoders remain unchanged. We design prompts for fine-tuning, outlined in Listings 2 and 3, that closely align with the format proposed in (Lei et al., 2023).

```
You are expert of multimodal emotion
classification and emotion cause
recognition.

The following is a conversation that
involves several speakers.

Here is a conversation that is described
in several fragments and includes
subtitles, video, and audio:

Utterance_1
<Speaker Name>: <Speaker Text>
Video: <Video>
Audio: <Audio>
...

Select the emotion label of each
utterance from <neutral, surprise,
fear, sadness, joy, anger, disgust>
and predict the ids of utterances that
caused this emotion.
```

Listing 2: Instruction format for the fine-tuning stage

```
Utterance_1
Emotion: <Emotion>
Causes: 1
...
```

Listing 3: Response format for the fine-tuning stage

4 Experiments

For the experiments described below, we use the Emotion-Cause-in-Friends (ECF) dataset (the official train part), which is divided into train, validation, and test sets in accordance with the proportion 8:1:1. We used train part due to test split is not officially available for extensive experiments.

4.1 Zero-shot and Few-shot

In the zero-shot experiments, the model exhibits a loss of ability to follow general instructions and ceases responding to the guidelines provided, instead demonstrating a tendency for a detailed description of the events observed in the video. Examples of this behavior are visible in the experimental data presented in Appendix B. In few-shot experiments with Video-LLaMA, we observe the same pattern.

4.2 Individual Embeddings

Metrics. In evaluating the model's performance on the emotion classification subtask, we utilize two principal metrics: the macro F1-score, which provides a balanced measure of precision and recall across all classes, and Accuracy, reflecting the overall proportion of correctly identified instances. For the causal classification subtask, we similarly measure performance using the binary F1-score, which is tailored to binary classification problems, alongside Accuracy to determine the proportion of true results in the dataset.

Training configuration. Each training session is run in 50 epochs. For emotion classification, 32 utterances are used as one batch. For cause classification, one batch describe one conversation and an accumulation of 6 batches for gradient optimization is used.

To address the challenges presented by MECAC, our approach encompassed two distinct training schemas: joint and separate training for the dual classification objectives, namely emotions and causes. The joint training final loss function was composed as a linear combination of the individual losses from both classification heads as in MTL systems (Karpov and Konovalov, 2023).

Initial observations from the joint training indicated that the combination of loss functions from the emotion and cause components was instrumental in enhancing the model's generalization capabilities. However, this joint strategy appeared to reach a plateau, failing to deliver the maximum attainable performance in the later stages of training.

Further experimentation yielded additional insights, particularly in the domain of model convergence. For the emotion classification task, the MLP head emerged as the superior architecture, leading to the most optimal model convergence. Conversely, the BiLSTM head demonstrated a marked advantage in the cause classification domain.

Also, as mentioned above for the training of cause classification it is suggested to perform balancing of the loss function. In practice, the assumption to mitigate class imbalance has proven to be highly significant. According to the experimental results, the best convergence was provided by the use of a constant weight coefficient, notably a value of 3, to give greater emphasis to predictions of the minor classes within the loss function.

Prompt optimization. We evaluate three distinct prompt configurations to derive embeddings for each modality under consideration. The experimental results reinforce the notion of textual content as a leading modality, with the third prompt configuration demonstrating particular efficacy. Accordingly, the tables in Appendix C present the training results as evaluated on the test subset, including all combinations of the prompts applied to the audio and video modalities.

Modality impact. To evaluate the contribution of each modality to the overall effectiveness of the classification tasks, we conducted several experiments. The validation results are depicted in Figure 1, and the test results confirm the observed trend. The text modality emerges as the most influential, exerting the greatest effect on the model's predictive accuracy. In a secondary position, the audio modality is found to have a considerable impact, albeit less than that of text. The video modality, while still contributing to the overall model performance, is observed to have the least influence among the three.

The leading role of textual modality can also be substantiated by the information provided by the ECF authors, who state that approximately 8% of the emotion causes in the dataset are the events mainly reflected in the acoustic or visual modalities. It's also important to note the least valuable modality in these experiments: the visual modality. We suppose that this is justified by the lack of confidence of the models in visual feature space, which, most likely, can be eliminated by fine-tuning of the visual encoding branch.

4.3 Fine-tuning

In the fine-tuning phase of our experiments, we employ LoRA technique to fine-tune the parameters of the language model component, specifically, the 4-bit quantized Llama-2 7b model, it's selected due to resource constraints. We configure LoRA with an alpha value of 16 and a low-rank factor of 8, while a dropout rate of 0.1 is utilized to prevent overfitting. Our method focuses on selectively adapting only the self-attention projection modules within the transformer architecture. This refines the model's focus on salient features for the tasks at hand without necessitating a comprehensive retraining of the entire network. Training batches are set to a size of 4, and the fine-tuning process is conducted over a single epoch, covering the full training dataset.

The fine-tuning strategy yields notable improvements in the model's performance across two distinct classification tasks. For the emotion classification task, the model achieves a macro F1-score of 0.6500 and an Accuracy of 0.7412. This represents a significant enhancement in the model's ability to discern and categorize emotional content within the input data accurately. In the causal classification task, the model demonstrates a binary F1-score of 0.3824 and an Accuracy of 0.9220. While the binary F1-score appears modest, the high Accuracy underscores the model's effectiveness in identifying causal relationships within the tested dataset.

5 Conclusion

In this paper, we conduct an analysis of the MLLM Video-LLaMA with an emphasis on its emotion understanding capabilities in MECAC. Our experiments show that multimodal models, in their current iteration, exhibit limitations in deciphering emotional states under zero-shot and few-shot modes.

To enhance the capabilities of such models in emotion understanding, our findings indicate that task-specific dataset fine-tuning is an essential step. Despite the challenges observed, the raw embeddings generated by the Video-LLaMA model show promising potential as a foundation for improving emotion recognition performance.

The implications of this research highlight the necessity for continued development and refinement of multimodal learning frameworks. Future work may concentrate on expanding the diversity of the datasets used for fine-tuning to include a broader spectrum of emotional expressions and cultural contexts. This could mitigate existing biases and enhance the model's generalizability across various demographics and scenarios. Moreover, incorporating advanced techniques such as transfer learning and domain adaptation could further enhance the model's proficiency in interpreting nuanced emotional states (Chizhikova et al., 2023).

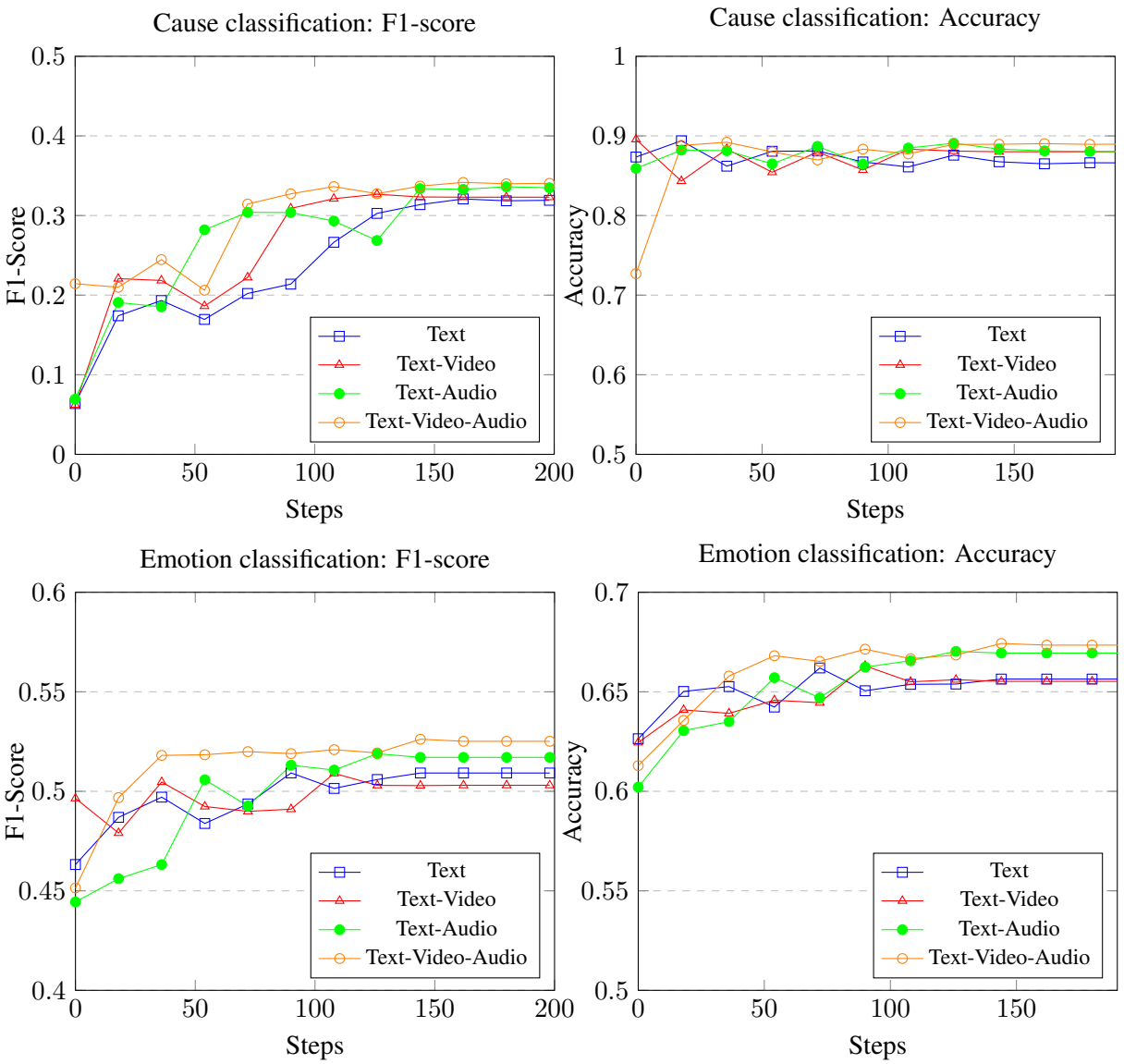

Figure 1: Impact of different modalities on the classification tasks

References

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. 2022. Flamingo: a visual language model for few-shot learning.

Ajwa Aslam, Allah Bux Sargano, and Zulfiqar Habib. 2023. Attention-based multimodal sentiment analysis and emotion recognition using deep neural networks. *Applied Soft Computing*, 144:110494.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023. Qwen technical report.

Anastasia Chizhikova, Vasily Konovalov, and Mikhail Burtsev. 2023. Multilingual case-insensitive named entity recognition. In *Advances in Neural Computation, Machine Learning, and Cognitive Research VI*, pages 448–454, Cham. Springer International Publishing.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang,

Boyang Li, Pascale Fung, and Steven Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning.

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.

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.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b.

Abhinav Joshi, Ashwani Bhat, Ayush Jain, Atin Vikram Singh, and Ashutosh Modi. 2022. Cogmen: Contextualized gnn based multimodal emotion recognition.

Dmitry Karpov and Vasily Konovalov. 2023. Knowledge transfer between tasks and languages in the multi-task encoder-agnostic transformer-based models. In *Computational Linguistics and Intellectual Technologies*, volume 2023.

Vasily Konovalov, Oren Melamud, Ron Artstein, and Ido Dagan. 2016. Collecting Better Training Data using Biased Agent Policies in Negotiation Dialogues. In *Proceedings of WOCHAT, the Second Workshop on Chatbots and Conversational Agent Technologies*, Los Angeles. Zerotype.

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.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023a. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models.

KunChang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. 2024a. Videochat: Chat-centric video understanding.

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, Limin Wang, and Yu Qiao. 2024b. Mvbench: A comprehensive multi-modal video understanding benchmark.

Wei Li, Yang Li, Vlad Pandelea, Mengshi Ge, Luyao Zhu, and Erik Cambria. 2023b. Ecpec: Emotion-cause pair extraction in conversations. *IEEE Transactions on Affective Computing*, 14(3):1754–1765.

Zheng Lian, Licai Sun, Mingyu Xu, Haiyang Sun, Ke Xu, Zhuofan Wen, Shun Chen, Bin Liu, and Jianhua Tao. 2023. Explainable multimodal emotion reasoning.

Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. 2023. Video-llava: Learning united visual representation by alignment before projection.

Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. 2023. Video-chatgpt: Towards detailed video understanding via large vision and language models.

Cam Van Thi Nguyen, Tuan Mai, Son The, Dang Kieu, and Duc-Trong Le. 2023. Conversation understanding using relational temporal graph neural networks with auxiliary cross-modality interaction. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Danil Tirskikh and Vasily Konovalov. 2023. Zero-shot ner via extractive question answering. In *Advances in Neural Computation, Machine Learning, and Cognitive Research VII*, pages 22–31, Cham. Springer Nature Switzerland.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem 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, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng 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.

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.

J. Wu, W. Gan, Z. Chen, S. Wan, and P. S. Yu. 2023. Multimodal large language models: A survey. In *2023 IEEE International Conference on Big Data (BigData)*, pages 2247–2256, Los Alamitos, CA, USA. IEEE Computer Society.

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.

Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, Chenliang Li, Yuanhong Xu, Hehong Chen, Junfeng Tian, Qian Qi, Ji Zhang, and Fei Huang. 2023. mplug-owl: Modularization empowers large language models with multimodality.

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

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.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models.

A Appendix

```
Video: <Conversation Video>
Speaker: <Speaker-1 Name>
Text: <Speaker-1 Text>
...
Speaker: <Speaker-i Name>
Text: <Speaker-i Text>

Predict last emotion of <Speaker Name-i>.
Choose emotion from this list: surprise, fear, sadness, neutral, joy, anger,
disgust.
```

Listing 4: Zero-shot prompt template

```
Video: <Conversation Video>
Speaker: <Speaker-1 Name>
Text: <Speaker-1 Text>
Emotion: <Speaker-1 Emotion>
...
Speaker: <Speaker-i Name>
Text: <Speaker-i Text>
Emotion: <Speaker-i Emotion>

Predict last emotion of <Speaker-i Name>.
Choose emotion from this list: surprise, fear, sadness, neutral, joy, anger,
disgust.
```

Listing 5: Few-shot prompt template

B Appendix

```
Speaker: Joey
Text: Let it go, Ross.
Speaker: Ross
Text: Yeah, well, you did not know Chi Chi.
Speaker: Monica
Text: Do you all promise?
Speaker: All
Text: Yeah! We promise! We will be good!
Emotion: _____

Describe the last emotion of all. Choose from the list: surprise, fear, sadness,
neutral, joy, anger, disgust.
```

Listing 6: Prompt for checking zero-shot capabilities

```
First, the video features a man, woman, and child sitting in a living room with a
green glass table, couch, and a flower vase. The mother is holding a blue blanket
while the man is eating a bowl of cereal. Next, the video switches to a shot of the
 mother holding a yellow flower while the daughter is holding a yellow tube and
talking on her phone. The mother is also seen holding a yellow flower in a
different shot.

The video then shows a couple holding a cake in a kitchen, with a wooden table and
chairs. The couple is standing in front of a counter with a white bowl and a white
ceramic plate. Then, the video shows a man, woman, and child sitting in a living
room with a flower vase on a green glass table, a couch, and a chair. The man and
woman are eating cereal while the child is holding a yellow tube and talking on the
 phone.

The video then shows a shot of the mother, daughter, and son sitting in a living
room with a green glass table, couch, and a chair. The mother is holding a blue
blanket while the son is eating cereal. Finally, the video features the mother
holding a blue blanket while the daughter is holding a yellow tube and talking on
her phone.
```

Listing 7: Model response for zero-shot prompt

C Appendix

Prompt format	macro F1-score	Accuracy
Audio-1 Video-1	0.4983	0.6554
Audio-1 Video-2	0.4992	0.6560
Audio-1 Video-3	0.5057	0.6568
Audio-2 Video-1	0.5102	0.6703
Audio-2 Video-2	**0.5251**	**0.6734**
Audio-2 Video-3	0.5131	0.6720
Audio-3 Video-1	0.5010	0.6566
Audio-3 Video-2	0.5105	0.6541
Audio-3 Video-3	0.5078	0.6575

Table 2: Prompt optimization results for emotion classification, where Modality-i is an i-th prompt option

Prompt format	F1-score	Accuracy
Audio-1 Video-1	**0.3505**	**0.8898**
Audio-1 Video-2	0.3496	0.8872
Audio-1 Video-3	0.3494	0.8850
Audio-2 Video-1	0.3480	0.8743
Audio-2 Video-2	0.3327	0.8735
Audio-2 Video-3	0.3194	0.8755
Audio-3 Video-1	0.3360	0.8806
Audio-3 Video-2	0.3325	0.8739
Audio-3 Video-3	0.3184	0.8799

Table 3: Prompt optimization results for cause classification, where Modality-i is an i-th prompt option