Explainable Multimodal Emotion Reasoning

Zheng Lian[1], **Licai Sun**[1,2], **Mingyu Xu**[1,2], **Haiyang Sun**[1,2], **Ke Xu**[1,2], **Zhuofan Wen**[1,2], **Shun Chen**[1,2], **Bin Liu**[1], **Jianhua Tao**[3]
[1]Institute of Automation, Chinese Academy of Sciences
[2]School of Artificial Intelligence, University of Chinese Academy of Sciences
[3]Department of Automation, Tsinghua University
lianzheng2016@ia.ac.cn

Abstract

Multimodal emotion recognition is an active research topic in artificial intelligence. Its primary objective is to integrate multi-modalities (such as acoustic, visual, and lexical clues) to identify human emotional states. Current works generally assume accurate emotion labels for benchmark datasets and focus on developing more effective architectures. But due to the inherent subjectivity of emotions, existing datasets often lack high annotation consistency, resulting in potentially inaccurate labels. Consequently, models built on these datasets may struggle to meet the demands of practical applications. To address this issue, it is crucial to enhance the reliability of emotion annotations. In this paper, we propose a novel task called "**Explainable Multimodal Emotion Reasoning (EMER)**". In contrast to previous works that primarily focus on predicting emotions, EMER takes a step further by providing explanations for these predictions. The prediction is considered correct as long as the reasoning process behind the predicted emotion is plausible. This paper presents our initial efforts on EMER, where we introduce a benchmark dataset, establish baseline models, and define evaluation metrics. Meanwhile, we observe the necessity of integrating multi-faceted capabilities to deal with EMER. Therefore, we propose the first multimodal large language model (LLM) in affective computing, called **AffectGPT**. We aim to tackle the long-standing challenge of label ambiguity and chart a path toward more reliable techniques. Furthermore, EMER offers an opportunity to evaluate the audio-video-text understanding capabilities of recent multimodal LLM. To facilitate further research, we make the code and data available at: https://github.com/zeroQiaoba/AffectGPT.

1 Introduction

Multimodal emotion recognition has experienced rapid development in recent years [1, 2]. Current works predominantly revolve around two aspects: the collection of larger and more realistic datasets [3, 4] and the development of more effective architectures [5, 6]. Despite promising progress, multimodal emotion recognition suffers from label ambiguity. It arises due to the inherent subjectivity of emotions, where different annotators may assign distinct labels to the same video. Label ambiguity results in potentially inaccurate labels of existing datasets, bringing obstacles to the systems developed on these datasets to meet high reliability in practical applications.

Current works primarily focus on increasing the number of annotators and using the majority vote to determine the most relevant emotion labels [7, 8]. Although this approach enhances the overall reliability of annotations, it may exclude correct but non-dominant emotion labels, limiting the model's ability to describe subtle emotions. In a cross-corpus setup, label ambiguity becomes more pronounced due to non-overlapping annotators. Consequently, combining multiple datasets generally fails to yield improved performance.

Preprint. Under review.

To address the problem of label ambiguity, the primary goal is to enhance the reliability of emotion annotations. In this paper, we change the existing emotion recognition framework by introducing a novel task called "Explainable Multimodal Emotional Reasoning (EMER)". Unlike previous works that primarily focus on predicting emotions, EMER takes a step further by providing detailed explanations for these predictions. The prediction is considered correct as long as the reasoning process behind the predicted emotion is plausible. To facilitate research in this area, we establish an initial dataset, baselines, and evaluation metrics. To the best of our knowledge, this is the first work to address multimodal emotion recognition in such an explainable manner. Meanwhile, we observe the necessity of integrating multi-faceted capabilities to deal with EMER. Therefore, we propose the first multimodal large language model (LLM) in affective computing, called AffectGPT.

Furthermore, researchers have started with initial attempts to extend large language models (LLMs) to multi-modalities [9–11]. However, there is an absence of benchmark datasets for evaluating the audio-text-video understanding capabilities. Since EMER focuses on capturing emotional clues in multi-modalities, it can be regarded as a fundamental task for multimodal LLMs. The main contributions of this paper can be summarized as follows:

- To enhance the reliability of emotion annotations, we introduce a new task called EMER. Unlike prior works that focus on predicting emotions, EMER further provides detailed explanations of these predictions.
- To facilitate research in this task, we construct an initial dataset, develop baselines, and define evaluation metrics. To address EMER, we further propose the first multimodal LLM in affective computing, called AffectGPT.
- Apart from enhancing the reliability of emotion predictions, this task can serve as a foundational task for evaluating the audio-video-text understanding ability of multimodal LLMs.

2 Dataset

In EMER, we should annotate emotion labels and provide evidence for these annotations. To construct the initial dataset, we select samples from the large-scale video emotion dataset, MER2023 [8]. Due to the high annotation cost, we randomly select 100 non-neutral samples to form the initial dataset. In the future, we plan to explore approaches to reduce annotation costs and increase the dataset size. The annotation process involves the following steps:

Clue Annotation. We employ six annotators and randomly assign three annotators to each video clip. Each annotator labels emotion clues from four aspects: 1) facial expressions and body movements; 2) tone and intonation; 3) speech content; 4) video content, environment, and other clues.

Clue Summarization. For each sample, three annotators provide clues from four aspects. To summarize all clues, we exploit ChatGPT [12] and use the prompt in Figure 1. However, we still observe some repeated expressions in the generated results. Therefore, we manually check and refine the outputs.

Emotion Summarization. In this step, we use ChatGPT to infer emotional states from the summarized clues, as this strategy provides more subtle emotions than the original labels in MER2023. However, we observe some unreliable emotions in the output. To address this issue, we use the few-shot ChatGPT for emotion summarization. The prompt is shown in Figure 2.

Combination of Clues and Emotions. We combine emotions and clues into one paragraph: *With the assistance of these "clues", we can infer the character's emotional state as "emotions"*. Then, we manually evaluate the plausibility of this reasoning process and get the final description.

3 Baselines

A straightforward approach for addressing EMER is to use multimodal LLMs, as these models are capable of handling various multimodal understanding tasks. Since emotion perception relies on temporal information, we only choose multimodal LLMs that support video inputs, including VideoChat [9], Video-ChatGPT [13], Video-LLaMA [10], PandaGPT [11], and Valley [14].

The basic idea behind multimodal LLMs is to align pre-trained models of other modalities to textual LLMs. For example, VideoChat and Video-LLaMA use Q-Former in BLIP-2 [15] to map visual

```
Multi-paragraph descriptions of a video is given below. Please summarize these descriptions as follows:
1. Please unify the subject of multiple paragraphs of "Clue Description" into "he".
2. Please summarize the multiple paragraphs of "Clue Description", delete repeated words, phrases or sentences, and describe the final result in complete sentences.
3. Check punctuation.

Input:
"Clue Description 1": {clue1}
"Clue Description 2": {clue2}
"Clue Description 3": {clue3}
"Clue Description 4": {clue4}
…
"Clue Description N": {clueN}

Output
```

Figure 1: Prompt for summarizing clues.

```
Please summarize the person's emotional state:

Input: He looks happy but is actually anxious.
Output: anxious

Input: {prediction}
Output:
```

Figure 2: Prompt for summarizing emotions.

queries to the textual embedding space. PandaGPT employs ImageBind [16] to learn alignment between six modalities. Video-ChatGPT and Valley exploit CLIP [17] to obtain text-aligned visual features. After fine-tuning on instruction datasets, these models exhibit remarkable abilities to understand instructions and multimodal inputs.

In addition to supporting video inputs, PandaGPT and Video-LLaMA also support audio inputs. Thus, we further feed audio into these baselines. To integrate subtitle information in the video, we include subtitles as additional content in the prompt. For a fair comparison, all baselines use the same prompt:

Prompt: The subtitle of this video is <Subtitle> <Subtitle_Here> </Subtitle>. Now answer my question based on what you have heard, seen, and given subtitles. From what clues can we infer the person's emotional state? Please summarize the clues in a maximum of 100 words.

4 Evaluation Metrics

Automatic Evaluation We exploit ChatGPT to evaluate prediction results from three aspects: 1) the degree of overlap between emotion-related clues; 2) the degree of overlap between summarized emotional states; 3) the modality completeness of the reasoning process. For the first two metrics, scores range from 0 to 10 with higher scores indicating greater overlaps. For these metrics, we first summarize emotion-related clues (or emotional states) using Prompt 1 and then calculate the overlap using Prompt 2 (see Figure 3 and Figure 4). Meanwhile, models that can infer emotions from more modalities should get higher scores. Therefore, we evaluate the completeness of the reasoning process using the prompt in Figure 5.

Human Evaluation Besides automatic evaluation, we also manually evaluate prediction results. For each video, we employ five annotators to judge the plausibility of the reasoning process. Annotators have four choices: "completely wrong", "correct (small part)", "correct (most part)", and "completely correct". We map these choices to scores ranging from 1 to 4, with higher scores indicating better reasoning ability.

Prompt1:
Please extract the description in relation to the character's emotional state, and then further summarize these descriptions.

input: {prediction}
output:

Prompt2:
"True clues" and "Predicted clues" are given below. Please compute the overlap between "True clues" and "Predicted clues". The higher overlap, the higher score returned. Scores range from 0-10.

True clues: lifted eyebrows, smiling face
Predicted clues: lifted eyebrows
Score: 5

True clues: {gt_clue}
Predicted clues: {pred_clue}
Score:

Figure 3: Prompt for calculating the degree of overlap between emotion-related clues.

Prompt1:
Please summarize the person's emotional state:

Input: He looks happy but is actually anxious.
Output: anxious

Input: {prediction}
Output:

Prompt2:
"True Emotion" and "Predicted Emotion" are given below. Please compute the degree of overlap between "True Emotion" and "Predicted Emotion". The higher overlap, the higher score returned. Scores range from 0-10.

True Emotion: happy
Predicted Emotion: happiness
Score: 10

True Emotion: {gt_emo}
Predicted Emotion: {pred_emo}
Score:

Figure 4: Prompt for calculating the degree of overlap between emotional states.

5 AffectGPT

In this paper, we use the initial EMER dataset to train an audio-video-text aligned multimodal LLM, called AffectGPT. The main framework is drawn from Video-LLaMA with some modifications. (1) Video-LLaMA trains audio and video branches separately. We modify it to support audio-video-text aligned training. (2) In Video-LLaMA, the input and output formats of different instruction datasets are inconsistent. Therefore, we unify the input and output formats. For more details please refer to our code: https://github.com/zeroQiaoba/AffectGPT.

```
Please summarize how many modalities are covered by the input. You can choose from ['audio', 'visual', 'content’]:

input: He looks happy and the music makes me happy.
output: visual, audio

input: He looks happy and the music makes me happy. Meanwhile, he express his gratitude to the researchers.
output: visual, audio, content

input: {prediction}
output:
```

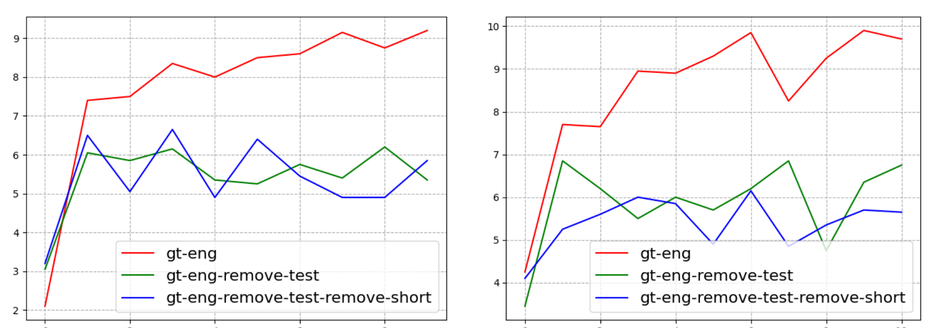
Figure 5: Prompt for evaluating the modality completeness.

We conduct some preliminary experiments to test the impact of different sample selection strategies. First, we randomly split the EMER dataset into 80 training samples and 20 testing samples. Then, we combine the EMER training set with three instruction datasets in Video-LLaMA (i.e., LLaVA [18], MiniGPT-4 [19], and VideoChat [9]). After instruction fine-tuning, we obtain the model “gt-eng-remove-test”. Meanwhile, we observe that longer videos generally have richer descriptions. Therefore, we remove short training samples (less than 2s) and obtain the model “gt-eng-remove-test-remove-short”. To investigate the upper-bound performance, we further test “gt-eng” that merges the training and testing sets during training.

In Figures 6∼7, we plot the curves of label overlap, clue overlap, and modality completeness with increasing training epochs. Compared with the initial model (see epoch=0), instruction fine-tuning with the EMER dataset generally brings performance improvement in emotional reasoning (see epoch>=1). Additionally, “gt-eng” can generate near-perfect results during training, reflecting the powerful fitting ability of LLMs.

In Figure 7, we observe some interesting phenomena. Most models can infer emotions from visual clues but few models can infer emotions from audio clues, which indicates that current multimodal LLMs mainly focus on the visual modality. To improve the multimodal understanding capability, it is necessary to consider more audio instruction datasets during training.

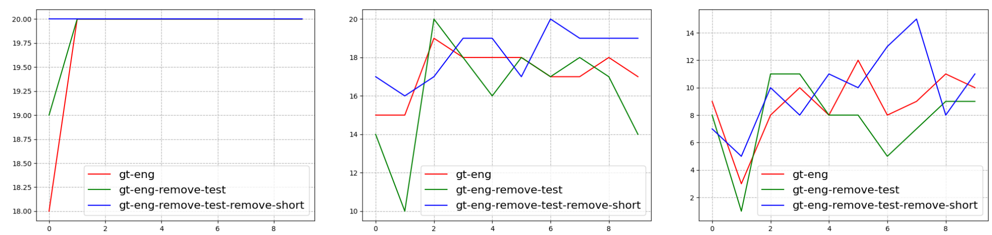
Furthermore, there is no significant difference between “gt-eng-remove-test” and “gt-eng-remove-test-remove-short” in terms of clue and label overlap but the latter has better modality completeness. The reason lies in that shorter videos generally have shorter emotion-related descriptions. After removing short samples, the model tends to generate longer descriptions and cover more modalities.



(a) clue overlap

(b) label overlap

Figure 6: Label overlap and clue overlap scores with increasing training epochs.



(a) visual (b) content (c) audio

Figure 7: Modality completeness with increasing training epochs.

6 Experimental Results and Discussion

6.1 Automatic Evaluation

In this section, we evaluate the performance of different baselines on the entire EMER dataset. We consider two versions of VideoChat: VideoChat-Text and VideoChat-Embed. The former uses vision models to convert visual data into the textual format, while the latter is an end-to-end model that aligns visual information to the textual embedding space. In Table 1, we observe a significant discrepancy between actual and predicted results, highlighting the limitations of existing multimodal LLMs in emotional reasoning. Meanwhile, the trends observed in these metrics exhibit a certain degree of similarity. Among all baselines, VideoChat-Text generally exhibits the worst performance, while Valley generally achieves the best performance. Notably, audio-considered baselines (such as PandaGPT and Video-LLaMA) do not show superior performance.

Table 1: ChatGPT-based evaluation results (scores range from 0∼10).

Models	Clue Overlap	Label Overlap
VideoChat-Text	6.41	4.00
Video-LLaMA	6.60	4.91
Video-ChatGPT	6.92	5.69
PandaGPT	7.17	5.49
VideoChat-Embed	7.14	5.65
Valley	7.21	5.78
Ground Truth	10.0	10.0

6.2 Human Evaluation

Due to the high cost of human evaluation, we only evaluate the performance of different methods on 20 testing samples. Besides prediction results, we also score the ground truth to eliminate the influence of human error and determine the upper-bound performance. For a fair comparison, we shuffle the order of evaluation, ensuring that annotators are unaware of which baseline or ground truth the description comes from. Table 2 shows the average scores of different baselines.

From Table 1 and Table 2, we observe some similarities between ChatGPT-based and human-based evaluation results. For example, VideoChat-Text consistently performs the worst, while Valley consistently achieves the best performance. Meanwhile, there are certain differences between these evaluation metrics. Therefore, ChatGPT-based scores can serve as a reference, but the main conclusions should come from human evaluation.

Meanwhile, we evaluate the effect of multi-model ensembles. Considering that there are some similarities between ChatGPT-based and human-based evaluation results, we further select the best predictions from all baselines based on clue overlap or emotion overlap. These models are denoted as "Baseline (Clue)" and "Baseline (Label)", respectively. In Table 2, we observe that this strategy can improve emotional reasoning performance, validating the advantage of the multi-model ensemble.

Table 2: Human evaluation results (scores range from 1$\sim$4).

Models	Human Evaluate
VideoChat-Text	1.73
Video-LLaMA	1.84
Video-ChatGPT	2.24
PandaGPT	2.09
VideoChat-Embed	1.88
Valley	2.45
Baseline (Clue)	2.60
Baseline (Label)	2.52
AffectGPT (Clue)	2.95
AffectGPT (Label)	3.19
Ground Truth	3.77

Furthermore, we reveal the performance of AffectGPT in Table 2. We select the best predictions from "gt-eng-remove-test" and "'gt-eng-remove-test-remove-short" across all epochs using clue overlap or emotion overlap. These models are denoted as "AffectGPT (Clue)" and "AffectGPT (Label)", respectively. Experimental results demonstrate that AffectGPT can achieve the highest score in emotional reasoning, fully verifying the effectiveness of our proposed strategy.

6.3 Qualitative Analysis

We randomly select two samples and visualize the prediction results of different baselines. In Figure 8, all baselines predict the person's emotional state as happy, despite the true label being angry. This misprediction occurs because existing baselines fail to comprehend acoustic clues (such as trembling voices and agitated tones) and facial clues (such as frowning). Furthermore, Video-LLaMA mistakenly identifies the presence of upbeat background music when there is no background music at all. We are impressed by the video description capabilities of Valley and Video-LLaMA. They successfully identify a man in a suit speaking into a microphone in front of wooden windows, even though these descriptions are unrelated to his emotional state.

In Figure 9, among all baselines, only PandaGPT and Video-ChatGPT accurately identify the person's emotional state as annoyed. However, Video-ChatGPT incorrectly identifies that the person is talking into a phone when he is not holding one. Although most of the baselines correctly recognize that the person is drinking water, this observation is unrelated to her emotional state.

These phenomena can be attributed to the fact that current multimodal LLMs are primarily trained on image caption datasets or video caption datasets, which focus on clothing, environment, action, etc., rather than facial-centric descriptions. Furthermore, these datasets usually ignore multimodal information, thereby restricting the audio-video-text understanding capabilities of multimodal LLMs trained on these datasets.

7 Conclusion

This paper introduces a novel task called EMER. Unlike traditional emotion recognition, our task goes beyond predicting emotional states by providing explanations for these predictions. By introducing this task, we aim to tackle the long-standing challenge of label ambiguity and enhance the reliability of emotion recognition systems. To facilitate research in this area, we construct an initial dataset, develop baselines, and define evaluation metrics. Experimental results demonstrate the difficulty of this task, as existing technologies struggle to achieve satisfactory results. Therefore, we propose AffectGPT which integrates multi-faceted capabilities to deal with EMER. Furthermore, EMER can also serve as a foundational task to evaluate the audio-video-text understanding abilities of current multimodal LLMs.

Chinese subtitle: 好，好，我的感受。我，我很开心，我很高兴，你满意了吗？
English subtitle: OK. OK. How did I feel? I, I was happy. I was very happy, now are you satisfied?

Ground Truth: In the video, he expressed his dissatisfaction and anger by asking rhetorically "Are you satisfied?" He spoke quickly, his tone was agitated, his voice was trembling, he stuttered, and his brows were furrowed. Therefore, it is presumed that his emotional state is anger.

PandaGPT: According to the subtitle "I was happy, I was very happy, now are you satisfied?", it can be inferred that this person's emotional state is happy and content. This person is expressing their positive emotions and is likely looking for validation or approval from the audience. This shows that the person is happy with the situation or outcome they are sharing, and they want to convey that happiness to the audience.

Valley: In a video, we can infer the emotional state of the person from their facial expressions and body language. They appear to be a man in a suit talking into a microphone. At one point, he's smiling and looking content, suggesting he might be happy or content. However, it is important to note that interpreting emotions from visual cues can be subjective, and we should consider other contextual information to more accurately confirm the person's emotional state.

VideoChat-Embed: Based on the video, we can infer that the person's emotional state is happy and content. The video shows the person smiling, a sign of happiness. The video's text indicates that the person is very happy and happy, while also feeling satisfied. Also, the person said they felt fulfilled and had no regrets. Overall, the person appears to be in a positive and happy state.

VideoChat-Text: Based on the subtitles provided, we can infer that the person is happy and happy. Using the word "very" emphasizes the intensity of these emotions. However, it is important to note that this inference is based only on subtitles, not any visual or audio cues in the video.

Video-ChatGPT: From the video it can be inferred that the person is happy and satisfied. He is seen smiling and laughing and the video shows him in a happy mood.

Video-LLaMA: The person is shown as a man in a suit speaking while standing in front of a wooden window. We see him fumbling and gesticulating as he speaks. He was happy and didn't show any signs of sadness or anxiety on his face. The background music was upbeat and made his speech sound even more enthusiastic. Based on these visual cues, we can infer that he is active and content.

Figure 8: Results for “sample_00006748.avi”.

In the future, we aim to reduce the annotation cost and increase the dataset size. Meanwhile, we will strive to design more effective baselines to improve the performance of EMER. We encourage more researchers to participate in this field to advance the practical applications of affective computing.

Chinese subtitle: 不要跟我扯上关系，我没有外公，也不需要他的礼物。
English subtitle: Don't get involved with me, I don't have a grandfather, and I don't need his gift.

Ground Truth: In the video, he expresses his displeasure with his grandfather by saying "Don't get involved with me, I don't have a grandfather, and I don't need his gifts", reflecting the tension between them. With a straight face, he spoke with his back to the other party, not wanting to pay attention to the other party. Therefore, it is presumed that his emotional state is anger.

PandaGPT: The people in this image appear to be in a negative emotional state as they hold a glass of water and look at the camera frowning. The subtitle "I don't have a grandfather and don't need his gifts" suggests that the person may feel disconnected or disconnected from family or people close to them. A frown and the negative tone of the caption indicate that the person is not in a positive or happy mood.

Valley: From the video, we can infer that the woman is probably in a calm or neutral emotional state. She is holding a bottle of water and drinking while looking at the camera. Additionally, she is wearing a gray sweater and necklace, and has brown hair. Her body language and facial expressions showed no obvious signs of strong emotions or critical reactions.

VideoChat-Embed: In the video, a woman drinks water. This indicates that she may be thirsty or need hydration. The woman's facial expressions and body language appear relaxed, without excess emotion. Based on these observations, it is likely that the woman is in a comfortable and relaxed state.

VideoChat-Text: Based on the provided video description, it is difficult to infer the person's emotional state, since there are no specific cues or behaviors that indicate a specific emotion. However, we can assume that this person is standing and drinking water in a professional setting, so he is probably calm and composed. It is important to note that emotions are subjective and may not always be accurately inferred from external cues.

Video-ChatGPT: From the video it can be inferred that the person may be annoyed and frustrated. He was seen talking to someone on the phone while the person on the other end was asking for a gift. The person on the other end of the line replied, "Don't get involved with me, I don't have a grandfather, and I don't need his gift." This means that the person on the other end of the line is trying to avoid this gift or is not interested in it. The person in the video may have been annoyed or frustrated by the request or the tone of the person on the phone.

Video-LLaMA: According to the video, the person is a woman wearing a suit jacket and holding a glass of water. She looks relaxed and at ease. The scene appears to be a modern office or workspace, with a desk and chair visible in the background. Based on the description and subtitles, we deduce that she doesn't have a grandfather and doesn't need his gifts. Therefore, it can be assumed that she is calm and content.

Figure 9: Results for “sample_00001421.avi”.

References

[1] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2):423–443, 2018.

[2] Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the 57th Conference of the Association for Computational Linguistics*, pages 6558–6569, 2019.

[3] Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Tensor fusion network for multimodal sentiment analysis. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1103–1114, 2017.

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

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

[6] Zheng Lian, Lan Chen, Licai Sun, Bin Liu, and Jianhua Tao. Gcnet: Graph completion network for incomplete multimodal learning in conversation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[7] Shan Li, Weihong Deng, and JunPing Du. Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2852–2861, 2017.

[8] Zheng Lian, Haiyang Sun, Licai Sun, Jinming Zhao, Ye Liu, Bin Liu, Jiangyan Yi, Meng Wang, Erik Cambria, Guoying Zhao, et al. Mer 2023: Multi-label learning, modality robustness, and semi-supervised learning. *arXiv preprint arXiv:2304.08981*, 2023.

[9] KunChang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. Videochat: Chat-centric video understanding. *arXiv preprint arXiv:2305.06355*, 2023.

[10] Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858*, 2023.

[11] Yixuan Su, Tian Lan, Huayang Li, Jialu Xu, Yan Wang, and Deng Cai. Pandagpt: One model to instruction-follow them all. *arXiv preprint arXiv:2305.16355*, 2023.

[12] OpenAI. Chatgpt. https://openai.com/blog/chatgpt/, 2023.

[13] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. *arXiv preprint arXiv:2306.05424*, 2023.

[14] Ruipu Luo, Ziwang Zhao, Min Yang, Junwei Dong, Minghui Qiu, Pengcheng Lu, Tao Wang, and Zhongyu Wei. Valley: Video assistant with large language model enhanced ability. *arXiv preprint arXiv:2306.07207*, 2023.

[15] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 1–13, 2023.

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

[17] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[18] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*, 2023.

[19] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023.